REPORT NUMBER 22

ROBUST, DISTRIBUTED, AND ADAPTIVE QUICKEST DETECTION PROCEDURES

R.W. CROW AND S.C. SCHWARTZ

INFORMATION SCIENCES AND SYSTEMS LABORATORY

Department of Electrical Engineering
Princeton University
Princeton, New Jersey 08544

June 1995

Prepared for

OFFICE OF NAVAL RESEARCH (Code 313SP)
Surveillance, Communications and Electronic Combat Division
Arlington, Virginia 22217
under Contract N00014-91-J-1144

S.C. Schwartz, Principal Investigator

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 22	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Robust, Distributed, and Adaptive Quickest Detection Procedures		5. TYPE OF REPORT & PERIOD COVERED Technical Report. Sept., 1991-June, 1994.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert W. Crow and Stuart Schwartz		8. CONTRACT OR GRANT NUMBER(s) N00014-91-J-1144
9. PERFORMING ORGANIZATION NAME AND ADDRESS Information Sciences & Systems Laboratory Dept. of Electrical Engineering Princeton University, Princeton, NJ 08544		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 411m006
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research (Code 313SP) Department of the Navy Arlington, VA 22217		12. REPORT DATE June 1995
		13. NUMBER OF PAGES 213
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Also submitted as the Ph.D. thesis of Robert W. Crow to the EE Department, Princeton University, Princeton, NJ 08544, August 1994

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

quickest detectors
robust quickest detection
distributed quickest detection
Page's test

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This dissertation focuses on sequential techniques for detecting a change, or disorder, in the statistics of a random process. First, the minimax robust quickest detector is derived for the case when the underlying noise models are only partially known. It is shown that when the robust processor is used, the minimax asymptotic performance measure is equal

Continued

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Continued

to the Kullback-Leibler divergence, and that the least favorable densities are those that minimize this quantity. The robust quickest detector is also determined for the weak signal case, and we show an equivalence between the performance measure, the classical efficacy, and Fisher's information. Performance curves are given to show the gain available when robustness is built into the procedure.

The robust quickest detector is also derived under mean and covariance uncertainty for a multivariate Gaussian noise process. It is shown that the robust processor is exactly the robust discrete-time matched filter, which has been studied previously. Expressions for the asymptotic performance are derived, and particular solutions are presented for several uncertainty classes. Performance curves are provided to illustrate the tradeoffs when there is a mismatch between the assumed and actual levels of uncertainty. The applicability of the robust procedure to non-Gaussian noise is also discussed.

Next, quickest detection procedures for the fusion processor of a distributed detection system are investigated. An optimal procedure is derived and compared to several alternative methods which are easier to implement in that they are recursive and require less computation. A simple method for choosing the thresholds of the local detectors is given, and a sensitivity analysis reveals that this choice results in overall system performance that is close to optimal. Lastly, performance curves are presented which illustrate the tradeoff between performance gain and channel bandwidth.

Finally, an adaptive procedure is proposed which is suitable for the disorder problem when a jump of unknown magnitude occurs in the mean of a random process. It is shown that this test exhibits asymptotic performance that is similar to the test which is optimal for known jump magnitude. The adaptive procedure is implemented to detect a change in the rate parameter of a Poisson process.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

This dissertation focuses on sequential techniques for detecting a change, or *disorder*, in the statistics of a random process. A disorder can be as simple as a shift in the mean from one constant to another, or as complex as a sudden change in the dynamic profile of multiple parameters. In either case, the overall goal is to determine *as soon as possible* that the change occurred, while at the same time minimizing the chance of falsely signalling the occurrence of a disorder in the absence of a change. In other words, we are seeking *quickest detection procedures.*

Many signal processing techniques assume that the parameters that characterize the data are either stationary or only slowly time-varying. However, there are numerous situations where this assumption does not hold. In such cases, quickest detection procedures can be used to signal the change so that some corrective action can be taken. Any area in which abrupt changes in the nature of a signal occur can potentially benefit from the use of quickest detection procedures. Many such examples can be found in the recent book by Basseville and Nikiforov [1].

The selection of a procedure for disorder detection is largely dependent on the particular application, as well as the amount of a priori information about the data.

In this thesis, three types of quickest detection problems are investigated: robust techniques which are suitable when the noise distributions are only partially known, quickest detection procedures designed for the fusion processor of a distributed detection system, and an adaptive procedure suitable for the case when the disorder is a jump in the mean of unknown magnitude. In each case, we are especially interested in seeking procedures that can be implemented recursively, making them suitable for on-line use.

1.2 Thesis Content

The body of this dissertation is divided into five chapters. Chapter 2 lays the foundation for the remainder of the thesis, as much of the notation and definitions are used in subsequent chapters. The disorder problem is presented formally, and previous work that is central to the field is reviewed. Many of the results of this work are represented by various performance curves, obtained either by direct computation or via Monte Carlo methods; the algorithms used to generate these plots are presented here.

Chapter 3 begins a study of robust quickest detectors. In many cases where quickest detection techniques would be desirable, the underlying statistical model may not be precisely known. Simply modelling the noise as Gaussian in this situation may result in the following problems: (1) the actual false alarm rate may differ significantly from the desired value, and (2) detectability may be sacrificed by simply increasing the decision threshold. In order to alleviate these problems, we derive the minimax robust detector based upon a lower bound on the asymptotic performance of Page's test. The robust procedure is derived for the epsilon-contaminated and total variation classes, both of which are useful in modelling real-world uncertainty. The performance of the robust procedure is compared to several nonparametric versions of Page's test,

which were studied in detail in [2]. The minimax robust procedure is also derived for the small-signal case. It is demonstrated that the robust procedure results in good performance over a wide range of noise distributions.

In Chapter 4, we consider the minimax robust quickest detection problem when the noise distribution is multidimensional Gaussian. It is shown that this problem is closely related to the previous work of Verdu and Poor [3] on minimax robust matched filtering, and that the solution to the latter problem can be used to solve the former. The robust procedure is derived for both signal and noise uncertainty, where the uncertainty is modelled as the deviation from some nominal parameters. The application of the robust quickest detector in multivariate non-Gaussian noise is also discussed.

In Chapter 5, we study the problem of determining as quickly as possible the occurrence of a disorder in a decentralized decision environment. Here, a number of sensors are used to monitor some phenomenon. The decision as to the presence or absence of a disorder is made at a central processor, or fusion center, based upon a summarized version of the sensor data. The processor receives a set of local binary decisions at regular intervals, where each decision indicates either "disorder present" or "no disorder present." The optimal procedure in the maximum likelihood sense is derived for this problem. For each set of local decisions, the fusion center must perform a search over all possible disorder times, a task which could become prohibitive when the local decisions are based on a large number of samples. It is shown that a small simplification can be made to eliminate the need for this search; this yields a suboptimal procedure which, nevertheless, exhibits performance nearly identical to the optimal version.

In perhaps the most significant contribution of this chapter, we propose a new simple method for choosing the thresholds of the local detectors based upon a lower bound on the asymptotic performance measure, which we derive for the distributed

detection problem. Direct optimization of this performance would require the solution of a set of constrained nonlinear equations. By comparison, the optimization of the lower bound on asymptotic performance is easy, requiring little computation. A sensitivity analysis reveals that the new method results in overall system performance which is close to optimal; this is particularly true when the false alarm rate is low, a condition which is desirable in many realistic scenarios. Each local decision is generated based upon a set of sequential sensor samples. In general, as the number of samples per local decision increases, both the required channel bandwidth to transmit the local decisions and the relative performance of the overall procedure decrease. We conclude the chapter by assessing this tradeoff. While perhaps contrary to intuition, it is shown that for the weak signal case, sending the local decision as frequently as possible does not result in the best performance.

In Chapter 6, we investigate the disorder problem when a jump of unknown magnitude occurs in the mean of a random process. An adaptive procedure is proposed that consists of two stages which operate sequentially: the first is a version of Page's test designed for a jump of minimum magnitude; the second is an adaptive version of the classical Wald sequential probability ratio test. The rationale behind such a test lies in the difficulty of reliably estimating the pre- or post-disorder means in the vicinity of the disorder time. For example, an estimate of the pre-disorder mean could likely become corrupted from samples from the post-disorder hypothesis, since the disorder time is unknown. The two-stage procedure provides a means to separate the two hypotheses (with some probability of error) so that the estimate of the mean after the disorder will be more reliable. It is shown that the adaptive test has similar asymptotic performance to the test which is optimal for known jump size. It also has the advantage of being recursive, more easily lending itself to on-line implementations. The procedure is implemented to detect a change in the rate parameter of a Poisson process. However, it is also applicable to other distributions.

Finally, the original contributions of this thesis are reviewed in Chapter 7.

1.3 References

[1] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] B. Broder, *Quickest Detection Procedures and Transient Signal Detection*, Ph.D. Thesis, Princeton University, 1990.

[3] S. Verdu and H. V. Poor, "Minimax robust discrete-time matched filters," *IEEE Trans. Commun.*, vol. COM-31, no. 2, Feb. 1983.

Chapter 2

Quickest Detection: A Review

The main body of this dissertation, Chapters 3 through 6, focuses on various problems that fall into the general category of quickest detection. The purpose of this chapter is to review some of the previous work in this area, as well as to introduce the definitions and notation that will be used throughout this thesis. Thus, this chapter will serve as a major reference for each of the subsequent chapters.

In Section 2.1, the disorder problem is first presented in a very general sense, and the goal of quickest detection is stated. Some of the assumptions that will be made throughout this thesis are also given. Section 2.2 introduces a procedure for on-line disorder detection known as Page's test; the optimality of this test is also discussed. In Section 2.3, we define the asymptotic performance measure for Page's test. This measure is a very useful quantity, and is a starting point for many of the results in this thesis. Section 2.4 presents several additional methods that may be used to compute the performance of quickest detection procedures. Finally, in Section 2.5, several extensions of quickest detection procedures to more complicated models are presented.

Much of the material in this chapter can be found in Chapter 2 of the Ph.D. thesis of Broder [5] and in the recent book by Basseville and Nikiforov [2].

2.1 The Disorder Problem

Consider a sequence of random variables $X_1, X_2, \ldots$, where random variable X_i has conditional density $f(X_i \mid \theta; \mathcal{X}_1^{i-1})$, $\mathcal{X}_1^j = X_1, \ldots, X_j$; here, θ is some scalar or vector quantity that parameterizes the conditional density. Suppose that $\theta = \theta_0$ for $i = 1, \ldots, m-1$, and $\theta = \theta_1$ for $i = m, m+1, \ldots$. In other words, the random variables undergo a *disorder* at time instant m, which is called the *disorder time*. The goal is to detect the change as soon as possible. [1] Thus, one wishes to detect a shift from hypothesis H_0 to hypothesis H_1, where

$$H_0 \;:\; X_i \sim f(X_i \mid \theta_0; \mathcal{X}_1^{i-1})$$

$$H_1 \;:\; X_i \sim f(X_i \mid \theta_1; \mathcal{X}_1^{i-1})$$

The above problem is phrased in terms of an *on-line* framework: the samples are received sequentially, and a decision regarding the occurrence of a disorder is made at each sample time. [2]

The change detection problem can alternatively be formulated using an *off-line* approach, where the decision is based on a finite "block" of samples $X_1, X_2, \ldots, X_n$. The problem here is to determine which of the hypotheses

$$\widetilde{H}_0 \;:\; X_i \sim f(X_i \mid \theta_0; \mathcal{X}_1^{i-1}), \quad \text{for } i = 1, \ldots, n$$

$$\widetilde{H}_1 \;:\; X_i \sim f(X_i \mid \theta_0; \mathcal{X}_1^{i-1}), \quad \text{for } i = 1, \ldots, m-1$$
$$X_i \sim f(X_i \mid \theta_1; \mathcal{X}_1^{i-1}), \quad \text{for } i = m, \ldots, n$$

holds. The off-line problem can be useful in situations where either the size, n, of the data window is small, or otherwise where there is ample computing power and memory for data storage. However, in most engineering applications, one is interested

[1] In this case, the disorder is a jump change in θ, but more general types of disorders may also be considered; some examples will be given in Section 2.5.

[2] The hypotheses can be written more generally as a change in the distribution $F(X_i \mid \theta; \mathcal{X}_1^{i-1})$. In this work, it is assumed throughout that the density functions exist, and so all of the expressions will be written in terms of $f(\cdot \mid \cdot)$.

in procedures that require little memory and may be implemented sequentially. In addition, it is sometimes the case that the disorder can be reliably detected using fewer than the n samples contained in the fixed block of data. Various techniques for off-line disorder detection are covered in [9]. However, in this thesis, we will consider only the on-line problem.

The following assumptions will be made throughout the thesis:

- **The disorder time m is unknown.**

  Two approaches are typically used in modelling the disorder time: the Bayesian approach, where m is modelled as a random variable, and the maximum likelihood (ML) approach, where m is taken to be unknown. The Bayesian approach was first investigated in [18], and is based on the assumption that the prior probability of the disorder time is known. On the other hand, the ML approach is more realistic when little is known about the disorder time, such as in situations where the waiting time before the disorder occurs is potentially very long. Examples include radar warning systems, where the threat (e.g., a missile) suddenly appears over the horizon, and communication link monitoring [16], where the channel characteristics may change suddenly due to some defect. In such instances, it may not be possible to accurately characterize the distribution of m.

- **The observations are independent**

  In this case, the on-line hypotheses become

$$H_0 \;:\; X_i \sim f(X_i \mid \theta_0) \triangleq f_0(X_i)$$

$$H_1 \;:\; X_i \sim f(X_i \mid \theta_1) \triangleq f_1(X_i)$$

  This assumption is made to simplify the problem, and will result in simpler algorithms. Examples of applications of quickest detection on uncorrelated data

include the detection of failures in linear systems via the monitoring of the innovations process [21], and the detection of changes in the drift in systems that can be modelled as a stochastic differential equation of the following form:

$$dX_t = \theta_t dt + dW_t$$

where

$$\theta_t = \begin{cases} \theta_0, & t < t_0 \\ \theta_1, & t \geq t_0 \end{cases}$$

and $\{W_t\}$ is a Weiner process [22]. [3] The latter can be used to model radar return, where the change in drift occurs when a target emerges. Other work investigates the problem of detecting disorders when the data is correlated. Examples where such problems arise are given in Section 2.5.

- **The disorder is a jump change in the mean**

  We consider the case where the parameter θ is simply the mean of the process, and that θ undergoes a one-time positive jump from θ_0 to $\theta_1 > \theta_0$. Not only does this assumption simplify the problem, but it is also a reasonable model for a large number of physical systems of interest. For example, the sudden failure of a device in a system may lead to a step change in the output. Also, sudden changes in spectral energy can be detected by testing for jumps in the coefficients of the energy spectral density [5]. Some examples involving more complicated changes are given in Section 2.5.

[3] X_t is uncorrelated when θ_t is scalar. For the vector case, the observables will likely be correlated.

2.2 Page's Test

In 1954, E. S. Page [15] introduced the following sequential procedure for detecting a shift from H_0 to H_1. Define the cumulative sum (CUSUM) statistic

$$T_j^k = \sum_{i=j}^{k} g(X_i) \tag{2.1}$$

where $g(x) = \log \frac{f_1(x)}{f_0(x)}$, with the convention that $T_j^k = 0$ if $j > k$. An alarm sounds (i.e., a disorder is declared) when the stopping time N occurs, where

$$N = \inf \left\{ n \mid T_1^n - \min_{0 \le k \le n} T_1^k > h \right\} \tag{2.2}$$

and $h > 0$ is some threshold. This procedure is commonly known as Page's test. Intuitively, one can see that this procedure terminates when the difference of the cumulative sum and its past minimum exceeds the threshold. It is easy to verify that when $g(\cdot)$ is the log-likelihood ratio, then $\mathrm{E}[g(x) \mid H_0] < 0 < \mathrm{E}[g(x) \mid H_1]$. [4] Therefore, T_1^n is seen to have a drift which is negative before the disorder and positive afterwards, and Page's procedure reacts to this change in drift. An example illustrating this point is shown in the upper plot of Figure 2.1.

Now consider the off-line version of the same problem discussed in Section 2.1. In particular, recall the definition of the hypotheses $\widetilde{H_0}$ and $\widetilde{H_1}$. The ML procedure for the off-line problem is to declare a disorder when the maximum of the log-likelihood ratio between $\widetilde{H_0}$ and $\widetilde{H_1}$ over all possible disorder times exceeds a threshold; in other

[4] This is done by using the relationships $1 - \frac{1}{x} \le \log x \le x - 1$, which are sometimes referred to as the "IT inequalities". We have

$$\mathrm{E}[g(x) \mid H_0] = \int \log \frac{f_1(x)}{f_0(x)} f_0(x) dx \le \int \left(\frac{f_1(x)}{f_0(x)} - 1 \right) f_0(x) = 0$$

and

$$\mathrm{E}[g(x) \mid H_1] = \int \log \frac{f_1(x)}{f_0(x)} f_1(x) dx \ge \int \left(1 - \frac{f_0(x)}{f_1(x)} \right) f_1(x) = 0$$

Equality holds only in the degenerate case where $f_0(x) \equiv f_1(x)$.

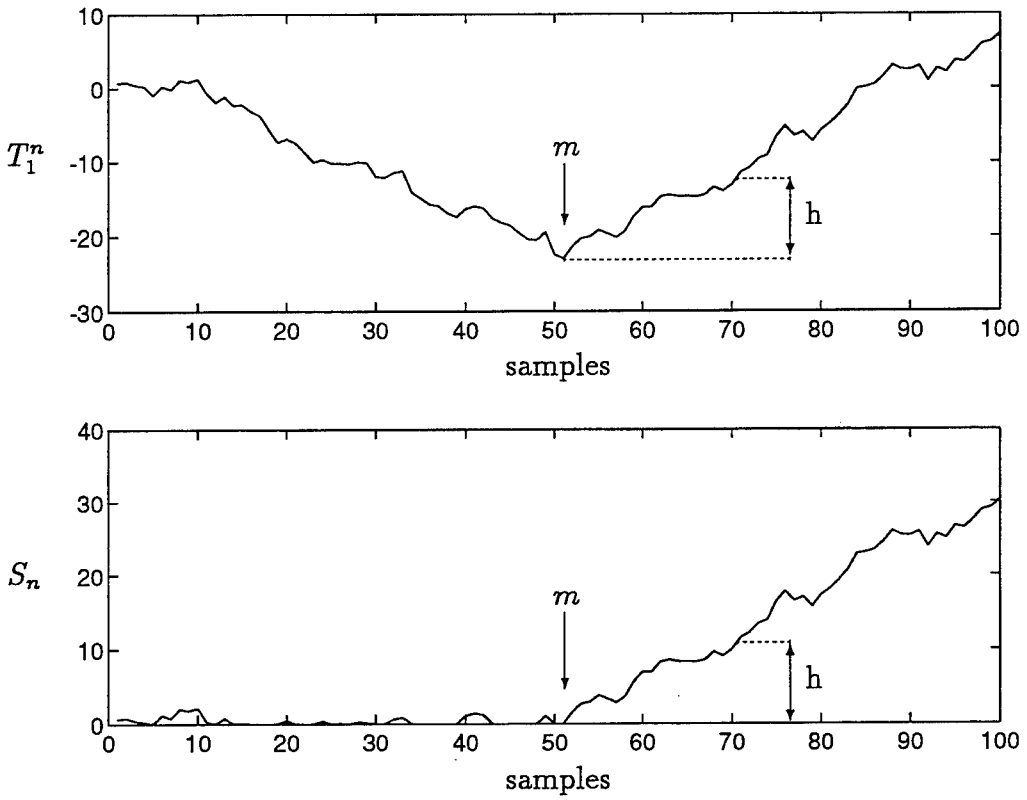

Figure 2.1: Two versions of Page's test.

words, declare a disorder in case

$$\max_{1\le m\le n} \log \frac{f(\mathcal{X}_1^n \mid \widetilde{H}_1)}{f(\mathcal{X}_1^n \mid \widetilde{H}_0)} = \max_{1\le m\le n} \sum_{i=m}^{n} \log \frac{f_1(X_i)}{f_0(X_i)} > h \tag{2.3}$$

It is not difficult to see that a sequential implementation of the above test, where a disorder is declared when the stopping time

$$N = \inf\left\{ n \mid \max_{1\le m\le n} \sum_{i=m}^{n} g(X_i) > h \right\}$$

occurs, is equivalent to the procedure in (2.2), where again g is the log-likelihood ratio. However, notice that with the off-line procedure, all past samples must be available at each iteration, while the sequential test only requires the storage of the past minimum of T_1^n. Thus, (2.2) can be interpreted as the on-line version of the ML procedure for detecting the change.

In [15], it is shown that an equivalent version of (2.2) is given by

$$N = \inf\{n \mid S_n > h\} \tag{2.4}$$

where S_n is generated by the recursion

$$S_n = \max\{S_{n-1} + g(X_n), 0\} \tag{2.5}$$

This version of Page's test will be used exclusively throughout the thesis; it is illustrated in the lower plot of Figure 2.1. Notice that S_n and T_1^n are exactly the same after the disorder time, and that both procedures react when the upward drift exceeds the threshold h. Like (2.2), this procedure is recursive and suitable for on-line applications. This version also has the advantage that the test statistic S_n always lies in the the interval $[0, h]$. On the other hand, T_1^n can potentially become very large in magnitude if the disorder occurs only after a long time; this could cause roundoff errors if it were necessary to quantize the samples using only a few bits. Also, observe that the test statistic in (2.5) can be interpreted as a repeated sequential probability ratio test (SPRT) with the continuation region $[0, h]$ in the following sense: if $S_n < 0$ (i.e., the lower boundary is crossed), the statistic is reset to zero and a new SPRT commences; if the upper boundary h is crossed, the test terminates and a disorder is declared. It will be shown later that this interpretation can be useful in computing the performance of Page's test. [5]

It turns out that Page's test implemented with $g(x) = \log \frac{f_1(x)}{f_0(x)}$ not only can be interpreted as a recursive ML procedure, but it is in fact the optimal procedure as explained below. Let N denote the stopping time of *any* procedure designed to detect the disorder. Define the following quantities:

$$T \triangleq \mathrm{E}_0 N$$

[5] In some cases the disorder is two-sided, such as a shift in the mean which may be either positive or negative. The approach here is to implement two Page's tests in parallel, one for each possible change direction, and declare a disorder when an alarm sounds in either test. In general, K parallel Page's tests may be used whenever there are K alternative hypotheses, H_k, $k = 1, \ldots, K$.

$$D \triangleq \sup_{k\geq 1} \text{ ess sup } \mathrm{E}_k\left[(N-k+1)^+ \mid X_1,\ldots,X_{k-1}\right] \triangleq \bar{\mathrm{E}}_1 N$$

where E_0 is the expectation under $\theta=\theta_0$, and E_k, for $k\geq 1$, is the expectation under the distribution of the observations when the change from θ_0 to θ_1 occurs at time k. T is called the *mean time between false alarms* (MFA), and D is the *worst expected delay* in detecting the disorder.

Let N denote the stopping time of Page's procedure using the log-likelihood. In [13], Lorden derives two key asymptotic results:

($\mathcal{R}1$) *Select the threshold h in (2.4) such that*

$$\mathrm{E}_0 N(\gamma) \geq \gamma$$

where we have indicated explicitly that N is a function of γ. Then as $\gamma\to\infty$,

$$\bar{\mathrm{E}}_1 N(\gamma) \approx \frac{\log\gamma}{I(\theta_1,\theta_0)}$$

where $I(\theta_1,\theta_0)$ is the Kullback-Leibler information number:

$$I(\theta_1,\theta_0) = \mathrm{E}\left[\log\frac{f_1(x)}{f_0(x)} \mid f_1\right]$$

($\mathcal{R}2$) *Let N' denote the stopping time of any other procedure, where*

$$\mathrm{E}_0 N'(\gamma) \geq \gamma$$

Then as $\gamma\to\infty$,

$$\bar{\mathrm{E}}_1 N'(\gamma) \gtrsim \frac{\log\gamma}{I(\theta_1,\theta_0)}$$

The first result characterizes the asymptotic performance of (2.4)-(2.5) as the (MFA) becomes large, a condition usually desired in real situations, since one would like the false alarms to be as infrequent as possible. The second result shows that no other test is asymptotically better than that in (2.4)-(2.5). In fact, it was later shown in

[14] and [17] that this procedure is also optimal in the non-asymptotic sense; that is, D is minimized for any fixed T.

Thus far, we have focused on the case where $g(\cdot)$ is the log-likelihood ratio, which results in the optimal version of Page's test. In some cases, though, the actual densities of the observations are not known, so the exact form of the log-likelihood ratio is not known and the optimal test cannot be implemented (for example, the exact values of θ_0 and θ_1 might be unknown). Therefore, it is also useful to consider the more general version of Page's test where $g(\cdot)$ is arbitrary.

In [5], nonparametric versions of Page's test are considered for the case where $g(\cdot)$ is the sign detector

$$g(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

and the dead-zone nonlinearity

$$g(x) = \begin{cases} -1, & x < -d \\ 0, & |x| \leq d \\ 1, & x > d \end{cases}$$

where $d > 0$. It is shown that these nonparametric quickest detectors are useful in cases where the underlying noise distributions are heavy-tailed. In Chapters 3 and 4 of this thesis, *robust* alternatives for quickest detection are investigated. Such techniques are useful when the noise is only partially characterized, and the goal is then to maximize the worst case performance. In this case, the nonlinearity $g(\cdot)$ is the solution of a minimax problem, and it turns out to be the log-likelihood of the least favorable distributions.

It will be necessary to characterize the performance of Page's test for arbitrary $g(\cdot)$. How this can be accomplished is the subject of the next two sections.

2.3 The Asymptotic Performance Measure

In designing a quickest detection procedure, one is interested in minimizing D for any operating point T, and this minimum occurs when g is the log-likelihood ratio between f_0 and f_1. Notice also that when this is the case, $(\mathcal{R}1)$ says that the worst expected delay is a logarithmic function of the MFA for large T. Therefore the performance of the optimal Page's test can be asymptotically characterized by the quantity

$$\eta = \lim_{T \to \infty} \frac{\log T}{D} \tag{2.6}$$

This quantity is called the *asymptotic performance measure* of Page's test, and $(\mathcal{R}1)$ implies that for the optimal Page test, $\eta = I(f_1, f_0)$. [6]

That η describes the asymptotic performance of Page's test can be seen by observing that $\frac{1}{\eta}$ is the slope of the plot of D versus $\log T$, as $T \to \infty$. In particular, for large T, we have the approximation

$$D \approx \frac{\log T}{\eta} \tag{2.7}$$

and so to minimize D, one needs to maximize η. Note that η is not a function of the threshold h, since the limit was taken as $h \to \infty$; this means that we have eliminated a variable from the optimization problem. However, one must make sure that the desired T is large enough so that (2.7) is valid. Luckily, in most practical problems, one is interested in designing procedures with few false alarms, resulting in large T.

As mentioned previously, Page's test can also be defined for arbitrary $g(x)$, and so it would also be useful to characterize the asymptotic performance for this case. Unfortunately, it is not clear how one would compute η for the generalized Page test, although we know from $(\mathcal{R}2)$ that $\eta \leq I(f_1, f_0)$. To address this problem, Broder [5] showed that the lower bound $\tilde{\eta} \leq \eta$ can be defined as follows:

$$\tilde{\eta} = \omega_0 \mathrm{E}\left\{g(x) \mid f_1\right\} \tag{2.8}$$

[6] Alternatively, the limit in (2.6) as $T \to \infty$ can be evaluated as either $h \to \infty$ or as $D \to \infty$, since any one of these implies the other two.

where ω_0 is the unique non-zero root of the moment generating function equality

$$\mathrm{E}\left\{e^{\omega_0 g(x)} \mid f_0\right\} = 1$$

An asymptotic upper bound on D can now be obtained as

$$D \approx \frac{\log T}{\eta} \leq \frac{\log T}{\tilde{\eta}}$$

Therefore, the upper bound on the worst expected delay can be minimized by selecting $\tilde{\eta}$ to be as large as possible.

It is shown in [5] that a sufficient condition for $\tilde{\eta}$ to be maximized (i.e., it equals η) is that g be the log-likelihood ratio. In this case, $\omega_0 = 1$, and $\tilde{\eta}$ directly reduces to the Kullback-Leibler divergence as expected. It is also shown that $\tilde{\eta}$ is invariant to changes in scale; thus, $\tilde{\eta} = \eta$ when $g(x) = C \log \frac{f_1(x)}{f_0(x)}$ for any $C > 0$. In Appendix A, we use variational calculus techniques to show the converse of this – that no other choice of $g(x)$ will make $\tilde{\eta} = \eta$; thus, $g(x) = C \log \frac{f_1(x)}{f_0(x)}$ is also a necessary condition.

The lower bound $\tilde{\eta}$ is useful for several reasons. First, it can be computed for any choice of $g(\cdot)$, enabling side-by-side comparisons of different tests. Second, as will be shown in later chapters, it is not difficult to compute. Finally, it enables us to obtain an upper bound on D for any (large) fixed T; thus, as will be seen in later chapters, a designer can use $\tilde{\eta}$ to quickly compute the approximate performance of a procedure which uses any nonlinearity $g(\cdot)$. We will use $\tilde{\eta}$ often in the next four chapters.

2.4 Methods of Performance Computation

We have seen that the performance of Page's test is characterized by the pair (T,D). Therefore, a natural way to compare several procedures is to compare the plots of D versus T for each one. It was shown in the previous section that the relationship between T and D can be approximated via the computation of $\tilde{\eta}$ as shown above. We now discuss how T and D can be obtained directly.

"Direct" Computation

Let $\mathcal{N}_z(\theta)$ denote the *average sample number* (ASN) of a CUSUM procedure whose initial score is z (i.e., $S_0 = z$). When the procedure begins, $S_0 = 0$, and so

$$T = \mathcal{N}_0(\theta_0)$$

Also notice that since $S_n \geq 0, \ \forall n$, the worst mean delay corresponds to the case where $S_{m-1} = 0$; therefore,

$$D = \mathcal{N}_0(\theta_1)$$

As stated earlier, Page's test can be viewed as a repeated application of a SPRT with lower boundary 0 and upper boundary h:

$$s_n = s_{n-1} + g(X_n), \quad s_0 = z$$

$$M = \inf\{n \mid s_n \notin [0, h]\}$$

Also define

$$\mathcal{M}_z(\theta) = \mathrm{E}[M \mid \theta]$$

and

$$\mathcal{P}_z(\theta) = \Pr\{s_M < 0 \mid \theta\}$$

$\mathcal{M}_z(\theta)$ is the ASN and $\mathcal{P}_z(\theta)$ is the *operating characteristic* of the SPRT, which is the probability that the SPRT will terminate at the lower boundary. It is not difficult to show that [15]:

$$\mathcal{N}_z(\theta) = \frac{\mathcal{M}_z(\theta)}{1 - \mathcal{P}_z(\theta)}$$

and so T and D can be written as

$$T = \mathcal{N}_0(\theta_0) = \frac{\mathcal{M}_0(\theta_0)}{1 - \mathcal{P}_0(\theta_0)}$$

and

$$D = \mathcal{N}_0(\theta_1) = \frac{\mathcal{M}_0(\theta_1)}{1 - \mathcal{P}_0(\theta_1)}$$

Define the transformation of random variables $Y = g(X)$, and let $\widehat{f}(y;\theta)$ and $\widehat{F}(y;\theta)$ denote the density and distribution of Y, respectively, conditioned on θ. The functions $\mathcal{M}_z(\theta)$ and $\mathcal{P}_z(\theta)$ satisfy the following Fredholm integral equations

$$\begin{aligned}\mathcal{P}_z(\theta) &= \widehat{F}(-z;\theta) + \int_0^h \mathcal{P}_y(\theta)\widehat{f}(y-z;\theta)dy \\ \mathcal{N}_z(\theta) &= 1 + \int_0^h \mathcal{N}_y(\theta)\widehat{f}(y-z;\theta)dy\end{aligned}$$

where $0 \leq z \leq h$. Unfortunately, no analytical solutions can be found for these equations. However, they can be approximated by discretizing the integral. The solution is determined by solving the system of linear equations

$$\mathcal{P}_{z_j}(\theta) = \widehat{F}(-z_j;\theta) + \sum_{k=1}^{K} w_k \mathcal{P}_{z_k}(\theta)\widehat{f}(z_k - z_j;\theta) \tag{2.9}$$

$$\mathcal{N}_{z_j}(\theta) = 1 + \sum_{k=1}^{K} w_k \mathcal{N}_{z_k}(\theta)\widehat{f}(z_k - z_j;\theta) \tag{2.10}$$

for $j = 1, \ldots, K$, where $0 \leq z_1 < z_2 < \ldots < z_{K-1} < z_K \leq h$, and where $w_1, \ldots, w_K$ is a set of weights chosen according to some rule. For example, when $\{z_k\}$ and $\{w_k\}$ are the roots and corresponding coefficients of the Legendre polynomial, (2.9)-(2.10) is called the Nyström approximation to the Fredholm equations of the second kind [6]. One could also use a simple rectangular approximation, which reduces the integrals to Riemann sums [5].

Markov Approximation

The ASN of Page's test can be expressed in another way. Let $r_i(n)$ denote the probability that stage n will be reached when $\theta = \theta_i$; that is,

$$r_i(n) = \Pr\{S_1, \ldots, S_{n-1} \in [0,h] \mid \theta_i\}$$

Now

$$\mathcal{N}_0(\theta_i) = \sum_{n=1}^{\infty} n(r_i(n) - r_i(n+1)) = \sum_{n=1}^{\infty} r_i(n) \tag{2.11}$$

Notice that $r_i(n) - r_i(n+1)$ is the probability that the test terminates in stage n.

The $r_i(j)$ can be computed using a finite state Markov chain approximation to Page's test, an approach introduced in [7]. The interval $[0, h]$ is divided into a total of p small bins of equal size, and each bin corresponds to a single state: specifically, state α_j corresponds to the subinterval $(x_{j-1}, x_j]$, where $x_j = \frac{jh}{p}$, $j = 1, \ldots, p$. The probability transition matrix, $\mathbf{Q}$, is formed, where element $Q_{i,j}$ denotes the probability of the test statistic S_n going from state α_i at time $n-1$ to state α_j at time n. Two additional states are also included. The first is the starting state, α_0, which corresponds to $S_n = 0$. The second is the terminal state, α^*, corresponding to the interval $(h, \infty]$; an alarm sounds whenever the terminal state is reached. It is shown in [7] that the structure of $\mathbf{Q}$ is

$$\mathbf{Q} = \left[\begin{array}{c|c} \mathbf{R} & (\mathbf{I}-\mathbf{R})\underline{1} \\ \hline \underline{0}^T & 1 \end{array}\right]$$

where $\underline{0}$ and $\underline{1}$ are $(p+1)$-dimensional column vectors of all zeros and ones, respectively. Separate transition matrices must be computed for $\theta = \theta_0$ and $\theta = \theta_1$: let these be $\mathbf{Q}_i$, $i = 0, 1$, respectively, with corresponding submatrices $\mathbf{R}_i$.

Let π_n denote the state probability vector at stage n:

$$\pi_n = [\Pr\{S_n \in \alpha_0\}, \ldots, \Pr\{S_n \in \alpha_p\}, \Pr\{S_n \in \alpha^*\}]$$

The successive state probabilities can be computed recursively [11] as

$$\pi_n = \pi_{n-1}\mathbf{Q}, \quad n = 1, 2, \ldots$$

The probability of reaching stage n is just the probability of not reaching the terminal state at time $n-1$. Thus,

$$r_0(n) = \pi_{n-1}\begin{bmatrix} \underline{1} \\ 0 \end{bmatrix} = \pi_0\mathbf{Q}^{n-1}\begin{bmatrix} \underline{1} \\ 0 \end{bmatrix}, \quad n = 1, 2, \ldots \tag{2.12}$$

where $\pi_0 = [1, 0, \ldots, 0]$ is the initial state probability vector. Equation (2.12) can be simplified to

$$r_0(n) = \pi_0' \mathbf{R}^{n-1} \underline{1}, \quad n = 1, 2, \ldots \tag{2.13}$$

where $\pi_0' = [1, 0, \ldots, 0]$ (dimension $p+1$). Note that $r_0(1) = 1$; that is, every test always requires at least one stage. Substituting (2.13) into (2.11), we have

$$\begin{aligned} \mathcal{N}_0(\theta_i) &= \sum_{n=1}^{\infty} \pi_0' \mathbf{R}_i^{n-1} \underline{1} \\ &= \pi_0' (\mathbf{I} - \mathbf{R}_i)^{-1} \underline{1} \end{aligned} \tag{2.14}$$

and so $T \approx \pi_0' (\mathbf{I} - \mathbf{R}_0)^{-1} \underline{1}$ and $D \approx \pi_0' (\mathbf{I} - \mathbf{R}_1)^{-1} \underline{1}$.

Monte Carlo Simulation

Finally, the ASN of Page's test can be obtained via Monte Carlo simulation in a straightforward manner. As mentioned above, T and D are just the ASN's of Page's test with initial score zero, which can be approximated by the average of the stopping times of K independent runs. Let N_k denote the stopping time of run k. An unbiased estimate of the ASN is

$$\widetilde{\mathcal{N}} = \frac{1}{K} \sum_{k=1}^{K} N_k$$

where $\widetilde{\mathcal{N}} \approx T$ when the samples are generated under $f(x; \theta_0)$ and $\widetilde{\mathcal{N}} \approx D$ when the samples are generated under $f(x; \theta_1)$. The Monte Carlo method will be particularly useful in Chapter 7, where we investigate an adaptive procedure which is not a version of Page's test, and for which the other methods cannot be applied directly.

2.5 Other Applications of Quickest Detection Procedures

To conclude this chapter, a brief survey of other areas where quickest detection procedures are applicable is given.

Closely related to the disorder problem discussed above is the problem of detecting transient signals. In this case, two shifts in the mean occur: from θ_0 to θ_1, and then back to θ_0. It is shown in [5] that the ML optimal procedure for the transient problem is again Page's test.

For the more general problem of correlated observations, a version of Page's test can be obtained by replacing the nonlinearity g by

$$g_n(X_n) = \log \frac{f_1(X_n \mid \mathcal{X}_i^{n-1})}{f_0(X_n \mid \mathcal{X}_i^{n-1})}$$

Thus, g is no longer memoryless, but now is a function of the past data. An auto regressive moving average (ARMA) model is commonly used to model correlated data. Here, the observations Y_n arise from the model

$$Y_n = \sum_{i=1}^{p} a_i Y_{k-i} + \sum_{j=0}^{q} b_j V_{k-j}$$

where V_k is a sequence of white Gaussian noise. The ARMA model is useful in spectrum modelling applications [10]. It can also be used to detect changes in spectral characteristics. Such changes correspond to a shift in the parameter set $\{a_1, \ldots, a_p, b_0, \ldots, b_q\}$. Examples include shifts in seismic, speech, and biomedical signals.

Most research in quickest detection has focused on the problem where the disorder is a shift from one stationary process to another. However, in some problems of practical interest, change may be time-varying. An example of this is the detection of sinusoidal signals for the purpose of carrier synchronization. In this case, the data

is not accurately modelled as a step change in the mean. In [4], Blostein derives a procedure that is suitable for detecting time-varying changes in the mean. Here, it is assumed that the mean before and after the disorder are known, and that the amplitude of the mean is at least approximately known. The procedure is similar to the time-varying version of Page's test, with the benefit that it can be implemented recursively. While this test is not optimal in the sense of Lorden [13], simulations reveal that the procedure works well for detecting sinusoidal signals of unknown amplitude in Gaussian noise. [7]

Even more difficult is the problem of detecting changes in systems where the statistics are not easily characterized, or where more than just the mean of the distributions is nonstationary. For example, in [1], the problem of detecting changes in geophysical systems is examined. These types of signals exhibit a high degree of nonstationarity (e.g. alternating segments of high and low variance) even when no disorder is present. Such signals can also arise in biomedical, speech, and image processing applications.

Work has also been done in detecting changes in the parameters of state-space systems. A typical example is a Kalman filtering application, where one wishes to track some phenomenon that is subject to sudden changes, such as a maneuvering target. In [21], a generalized likelihood ratio (GLR) approach is introduced to handle this problem. The presence of a disorder can be determined by monitoring the filter residual process: the residual is white Gaussian noise, with zero mean before the disorder and nonzero mean afterwards. When a disorder is detected, an estimate of the disorder magnitude is determined and used to adjust the model parameters; in essence, the model is bootstrapped for the new statistics. [8] The application of the GLR procedure to geophysical signals is discussed in [1]. A survey of failure detection

[7] Lorden's proof of optimality requires that the samples before and after the disorder time be independent and identically distributed.

[8] It is assumed that the system is observable so that any change in the state variables will show up in the residual signal.

in dynamic systems is given in [20].

2.6 Appendix

2.6.A The Log-likelihood Ratio is Necessary and Sufficient to Maximize $\tilde{\eta}$

Proposition 1: *A necessary and sufficient condition that $\tilde{\eta}$ is maximized is that*

$$g(x) = C \log \frac{f_1(x)}{f_0(x)}$$

for some $C > 0$.

Proof:

($\Leftarrow$)

To prove sufficiency, simply let $g(x) = C \log \frac{f_1(x)}{f_0(x)}$. We have

$$\begin{aligned} \tilde{\eta} &= \omega_0 C \int_{-\infty}^{\infty} \log \frac{f_1(x)}{f_0(x)} f_1(x) dx \\ &= \omega_0 C I(f_1, f_0) \end{aligned} \tag{A.1}$$

where ω_0 satisfies

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \exp\left\{\omega_0 C \log \frac{f_1(x)}{f_0(x)}\right\} f_0(x) dx \\ &= \int_{-\infty}^{\infty} \left[\frac{f_1(x)}{f_0(x)}\right]^{\omega_0 C} f_0(x) dx \end{aligned}$$

The latter implies $w_0 = C^{-1}$. Therefore, (A.1) becomes

$$\tilde{\eta} = I(f_1, f_0)$$

In addition, Lorden has shown [13] that optimal performance over all possible choices of $g(x)$ is

$$\eta = I(f_1, f_0)$$

and that this occurs when $g(x)$ is the log-likelihood ratio. Therefore, $\widetilde{\eta}$ is maximized.

($\Rightarrow$)

To show the converse, consider the following constrained optimization problem:

$$\textit{maximize over all } g: \qquad \widetilde{\eta} = \omega_0 \int_{-\infty}^{\infty} g(x) f_1(x) dx$$

$$\textit{subject to the constraint: } \int_{-\infty}^{\infty} \exp\{\omega_0 g(x)\} f_0(x) dx = 1$$

where ω_0 is any fixed positive real number. This is a so-called *isoperimetric* problem from variational calculus [8, 12, 19]. The solution is obtained by first incorporating the side constraint via the Lagrange multiplier method, and then applying the standard calculus of variations optimization procedure.

The goal is to determine the $g(x)$ for which the integral

$$\int_{-\infty}^{\infty} [\omega_0 g(x) f_1(x) + \lambda \exp\{\omega_0 g(x)\} f_0(x)] \, dx \tag{A.2}$$

is stationary. Here, λ is the Lagrange multiplier associated with the side constraint. Suppose that $\widehat{g}(x)$ is the function which maximizes $\widetilde{\eta}$, and consider the nonlinearity

$$g(x) = \widehat{g}(x) + \epsilon \cdot \delta g(x)$$

where $\delta g(x)$ is an arbitrary variation in the neighborhood of $\widehat{g}(x)$. Substituting this into (A.2), we have

$$K(\epsilon) = \int_{-\infty}^{\infty} \left\{ \omega_0 [\widehat{g}(x) + \epsilon \cdot \delta g(x)] f_1(x) + \lambda e^{\omega_0 [\widehat{g}(x) + \epsilon \cdot \delta g(x)]} f_0(x) \right\} dx$$

Now, a *necessary* condition to get a stationary point is

$$\left. \frac{dK(\epsilon)}{d\epsilon} \right|_{\epsilon=0} = 0$$

The derivative can be taken inside the integral, and thus we have

$$\left.\frac{dK(\epsilon)}{d\epsilon}\right|_{\epsilon=0} = \int_{-\infty}^{\infty} \left[\omega_0 \delta g(x) f_1(x) + \lambda \omega_0 \delta g(x) e^{\omega_0 \widehat{g}(x)} f_0(x)\right] dx = 0$$

Rearranging terms, we have

$$\omega_0 \int_{-\infty}^{\infty} \delta g(x) \left[f_1(x) + \lambda e^{\omega_0 \widehat{g}(x)} f_0(x)\right] dx = 0$$

In order for the equality to hold for arbitrary variations δg, the expression within brackets must be zero for *all* x. Therefore, the necessary condition becomes

$$f_1(x) + \lambda e^{\omega_0 \widehat{g}(x)} f_0(x) = 0$$

which can be rearranged to get

$$\frac{f_1(x)}{f_0(x)} = e^{\log(-\lambda) + \omega_0 \widehat{g}(x)}$$

and thus

$$\omega_0 \widehat{g}(x) = \log \frac{f_1(x)}{f_0(x)} - \log(-\lambda) \tag{A.3}$$

We can now use the equality constraint to determine λ. Observe that

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} \exp\{\omega_0 \widehat{g}(x)\} f_0(x) dx \\
&= \int_{-\infty}^{\infty} \exp\left\{\log \frac{f_1(x)}{f_0(x)} - \log(-\lambda)\right\} f_0(x) dx \\
&= \int_{-\infty}^{\infty} \frac{f_1(x)}{f_0(x)} \left(\frac{1}{-\lambda}\right) f_0(x) dx \\
&= -\frac{1}{\lambda} \int_{-\infty}^{\infty} f_1(x) dx = -\frac{1}{\lambda}
\end{aligned}$$

Therefore, $\lambda = -1$, and (A.3) becomes

$$\widehat{g}(x) = \frac{1}{\omega_0} \log \frac{f_1(x)}{f_0(x)}$$

which is just the log-likelihood ratio scaled by the factor $\frac{1}{\omega_0}$. Finally, ω_0 was chosen arbitrarily; therefore the optimal nonlinearity is the log-likelihood ratio scaled by any positive constant.

That $\widehat{g}(x)$ results in a maximum (rather than some other stationary point) can be seen by noting that

$$\begin{aligned}\left.\frac{d^2K(\epsilon)}{d\epsilon^2}\right|_{\epsilon=0} &= \left.\frac{d}{d\epsilon}\int_{-\infty}^{\infty}\left[\omega_0\delta g(x)f_1(x)+\lambda\omega_0\delta g(x)e^{\omega_0[\widehat{g}(x)+\epsilon\delta g(x)]}f_0(x)\right]dx\right|_{\epsilon=0} \\ &= \lambda\omega_0^2\int_{-\infty}^{\infty}\delta g^2(x)e^{\omega_0\widehat{g}(x)}f_0(x)dx \qquad \text{(A.4)}\end{aligned}$$

Clearly the integral is strictly positive and we have seen that $\lambda = -1$. Therefore,

$$\left.\frac{d^2K(\epsilon)}{d\epsilon^2}\right|_{\epsilon=0} < 0$$

and so the stationary point is in fact a maximum.

Also, it is easy to verify that when $g(x) = \widehat{g}(x)$, $\widetilde{\eta} = I(f_1, f_0)$. Therefore, $\widetilde{\eta} = \eta$. ∎

2.7 References

[1] M. Basseville and A. Benveniste, "Design and comparative study of some sequential jump detection algorithms for digital signals," *IEEE Trans. Acoust. Speech Sig. Proc.*, vol. ASSP-31, no. 3, June 1983.

[2] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[3] A. Benveniste, "Advanced methods of change detection: An overview," *Detection of Abrupt Changes in Signals and Dynamical Systems*, ed. Basseville, M. and Benveniste, A., Springer-Verlag, New York, 1986.

[4] S. D. Blostein, "Quickest detection of a time-varying change in distribution," *IEEE Trans. Info. Theory*, vol. 37, no. 4, July 1991.

[5] B. Broder, *Quickest Detection Procedures and Transient Signal Detection*, Ph.D. Thesis, Princeton University, 1990.

[6] I. N. Bronshtein and K. A. Semendyayev, *Handbook of Mathematics*, Van Nostrand Reinhold Co., New York, 1985.

[7] D. Brook and D. A. Evans, "An approach to the probability distribution of cusum run length," *Biometrika*, vol. 59, no. 3, pp. 539-549, 1972.

[8] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Interscience Publishers, Inc., New York, 1953.

[9] J. Deshayes and D. Picard, "Off-line statistical analysis of change-point models using nonparametric and likelihood methods," *Detection of Abrupt Changes in Signals and Dynamical Systems*, ed. Basseville, M. and Benveniste, A., Springer-Verlag, New York, 1986.

[10] S. Kay, "Spectral Estimation," *Advanced Topics in Signal Processing*, ed. J. S. Lim and A. V. Oppenheim, Prentice Hall, Englewood Cliffs, NJ, 1988.

[11] L. Kleinrock, *Queueing Systems, Vol. 1*, John Wiley & Sons, New York, 1975.

[12] H. A. Lauwerier, *Calculus of Variations in Mathematical Physics*, Mathematical Centre Tracts, Amsterdam, 1966.

[13] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897-1908, 1971.

[14] G. Moustakides, "Optimal procedures for detecting changes in distributions," *Ann. Statist.*, vol 14., no. 4, pp. 1379-1387, 1986.

[15] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100-114, 1954.

[16] P. Papantoni-Kazakos, "Algorithms for monitoring changes in the quality of communications links," *IEEE Trans. Commun.*, vol. COM-27, no. 4, April 1979.

[17] Y. Ritov, "Decision theoretic optimality of the CUSUM procedure," *Ann. Statist.*, vol. 18, no. 3, pp. 1464-1469, 1990.

[18] A. N. Shiryaev, "The problem of the most rapid detection of a disturbance in a stationary regime," *Sov. Math. Dokl.*, no. 2, pp. 795-799, 1961.

[19] J. B. Thomas, *An Introduction to Communication Theory and Systems*, Springer-Verlag, New York, 1988.

[20] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, vol. 12, pp. 601-611, 1976.

[21] A. S. Willsky and H. L. Jones, "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 108-112, February 1976.

[22] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1985.

Chapter 3

Robust Quickest Detection

3.1 Introduction

In the previous chapter, it was shown that optimal procedures for quickest detection exist when the noise distribution is known and the samples are independent. Unfortunately, in practice the true noise distribution is often not known precisely. This leads to two questions. First, suppose a procedure is optimal for a particular noise distribution. How sensitive is the performance of this procedure when the true noise distribution deviates from the assumed distribution? Second, if it is only known that the true distribution lies within some noise uncertainty class, what then is the optimal procedure? Both of these questions are addressed in this chapter.

The disorder problem for a shift in the mean with noise uncertainty is very similar to that for known noise characteristics stated in Chapter 2. As before, assume that a sequence of independent random variables $X_1, X_2, \ldots$ is observed sequentially. Let H_0 and H_1 define the hypotheses "no disorder is present" and "the disorder has occurred," respectively. At the *disorder time* m, a one-time shift in the mean from $-\theta$ to θ occurs, where $\theta > 0$. Let the noise distributions before and after the disorder be f_0 and f_1, respectively. If these distributions are known, we have seen in Chapter 2 that the optimal procedure is Page's test implemented using the log-likelihood.

In this chapter, however, it is assumed that, rather than having perfect knowledge of the noise characteristics, f_0 and f_1 are instead known only to lie within some noise classes $\mathcal{F}_0$ and $\mathcal{F}_1$. For this case, the disorder problem involves detecting as quickly as possible after time instant $i = m$ that a shift from hypothesis H_0 to hypothesis H_1 has occurred, where:

$$H_0: \quad X_i \sim f_0 \in \mathcal{F}_0, \quad i = 1, 2, \ldots, m-1$$

$$H_1: \quad X_i \sim f_1 \in \mathcal{F}_1, \quad i = m, m+1, \ldots$$

As before, the goal is to minimize the expected delay in detecting the disorder, D, subject to a lower bound on the mean time between false alarms (MFA), T. However, it is necessary to be more specific about what it means to achieve *optimal* performance when the noise pair (f_0, f_1) is known only to lie within some class $\mathcal{F}_0 \times \mathcal{F}_1$.

All of the noise uncertainty classes that will be considered here consist of a specific nominal distribution together with some type of allowable uncertainty; in particular, attention will be paid to the case of Gaussian nominals, due to their widespread applicability. One design option is to implement the procedure which is optimal for the nominal distribution and assume (hope!) that the performance will be similar in the case where the true noise deviates from this nominal. If the size of the noise classes is small, this design philosophy may prove satisfactory; on the other hand, if the noise classes are large, it is not clear what the outcome might be. At the very least, it would be nice to know what performance results when the true noise is not the nominal. A more desirable design scheme would be to optimize the performance over the entire uncertainty class. Unfortunately, a single procedure which is optimal for each distribution within the class does not usually exist. [1]

An alternative design methodology is to determine the test which optimizes the

[1] An exception to this, for example, is the noise class consisting of univariate Gaussian distributions whose variance may lie on some interval. In this case, Page's test implemented with $g(x) = x$ is optimal regardless of the true distribution; this follows from the invariance of $\bar{\eta}$ with respect to scale changes in $g(x)$, which was shown in Chapter 2.

performance for the *least favorable* noise distributions in $\mathcal{F}_0$ and $\mathcal{F}_1$; this is the well-known *minimax* design philosophy. [2] A disadvantage of this scheme is that the performance will be less than optimal when the noise is in fact generated by the nominal distributions; however, the worst case performance will be maximized. This is the key reason for considering robust procedures: not only can one get reasonable performance over the entire noise class, but the performance can also be guaranteed to be at least some minimal value.

In Section 3.2, the solution to the robust quickest detection problem is given. The approach involves applying the minimax criterion to the asymptotic performance measure introduced in Chapter 2. It is shown that there is a direct connection between robust quickest detection and robust hypothesis testing. As a result, many of the results from the latter can be applied to the present problem.

The exact forms of the robust quickest detector for two noise uncertainty classes, the *ε-contaminated* and *total variation* classes, are given in Section 3.3. Expressions for a lower bound on the asymptotic performance are computed for: *i)* the robust procedure, *ii)* the test which is optimal for the nominal distributions, and *iii)* two nonparametric alternatives. The computation is done for several members of each noise uncertainty class.

In Section 3.4, the lower bounds are evaluated for each of the noise/detector combinations over a range of parameters. First, it is shown that the bounds are actually good approximations to the true asymptotic performance. Next, the asymptotic performance is computed for a range of signal to noise ratios. A useful figure of merit, the robustness index, is also computed: this is a measure of the performance gain (or loss) in opting for the robust procedure over each of the others. The effect of the level of uncertainty assumed in the noise model is evaluated. The section concludes

[2] Technically, this should be called "maximin", since the performance here is measured by a *gain* function rather than a *loss* function. However, as is usually done, the term "minimax" will be used throughout with the true idea being clear from the context.

with an example illustrating the utility of the asymptotic performance measure in practical applications where high MFAs are desired.

In Section 3.5, the robust quickest detector for the weak signal case is determined. Again, there is a strong connection with robust hypothesis testing, and therefore some of the previous work can be exploited. The optimal robust detector is determined for the ε-contaminated noise class by applying the minimax criterion directly to the efficacy, which is proportional to the weak signal asymptotic performance. The robustness index in this case simply reduces to the asymptotic relative efficiency between the robust and alternative procedures. Finally, some performance curves are computed to illustrate the benefits of the robust procedure in weak signal applications.

3.2 Robust Asymptotic Performance

In Chapter 2, the *asymptotic performance measure* for Page's test

$$\eta = \lim_{T\to\infty} \frac{\log T}{D} \tag{3.1}$$

was defined, [3] along with the lower bound

$$\tilde{\eta} = \omega_0 \mathrm{E}\left[g(X) \mid H_1\right] \leq \eta$$

where ω_0 satisfies the moment generating function equality

$$\mathrm{E}\left[\exp\{\omega_0 g(X)\} \mid H_0\right] = 1$$

In most situations where quickest detection procedures are applicable, one is interested in procedures where false alarms occur infrequently, in other words, where T is large. In this case, we see from (3.1) that T and D can be related by

$$D \approx \frac{\log T}{\eta} \leq \frac{\log T}{\tilde{\eta}} \tag{3.2}$$

[3] Unless noted otherwise, "log" denotes the natural logarithm.

Since the ultimate goal is to minimize D for fixed T, then, for large T, an approximately equivalent strategy is to maximize $\widetilde{\eta}$. [4] Similarly, in order to obtain the robust quickest detector, the minimax criterion can be applied directly to $\widetilde{\eta}$.

We wish to maximize the asymptotic performance of Page's test for the *least favorable* noise distributions $(f_{0L}, f_{1L}) \in \mathcal{F}_0 \times \mathcal{F}_1$. Recall that Page's test involves the computation of a cumulative sum test statistic, where each sample is processed by a nonlinearity g, and that an alarm sounds when this statistic exceeds some threshold h. The asymptotic performance measure describes the limiting behavior of the performance as $T \to \infty$, which is also equivalent to $h \to \infty$ (or as $D \to \infty$). Therefore, in designing a version of Page's test for practical (i.e., for large T) applications, one need only consider the choice of g.

Let $\mathcal{G}$ denote the set of all memoryless functions g. The direct minimax problem is

$$\max_{g \in \mathcal{G}} \min_{(f_0, f_1) \in \mathcal{F}_0 \times \mathcal{F}_1} \eta(g; f_0, f_1) = \eta(g_R; f_{0L}, f_{1L}) \tag{3.3}$$

In order to make it easier to solve this problem, we would like to find a *saddlepoint solution* of (3.3); that is, we would like to determine some $(g_R, (f_{0L}, f_{1L}))$ that satisfies

$$\max_{g \in \mathcal{G}} \eta(g; f_{0L}, f_{1L}) = \eta(g_R; f_{0L}, f_{1L}) = \min_{(f_0, f_1) \in \mathcal{F}_0 \times \mathcal{F}_1} \eta(g_R; f_0, f_1) \tag{3.4}$$

This allows the maximization and minimization to be performed separately, rather than jointly, thus simplifying things considerably. The following proposition establishes that a saddlepoint does exist for this problem:

Proposition 2: *There exists a saddlepoint solution, $(g_R, (f_{0L}, f_{1L}))$, of (3.3).*

[4] Although $\widetilde{\eta}$ is a lower bound on η, it will be seen later in this chapter that $\widetilde{\eta} \approx \eta$ in most cases of interest.

Proof:

In [7], Lorden proved that no procedure has better asymptotic performance than Page's procedure implemented using the log-likelihood ratio. Therefore, for any pair (f_0, f_1), we have

$$\max_{g \in \mathcal{G}} \eta(g; f_0, f_1) = \eta(g^*; f_0, f_1)$$

where $g^*(x) = \log \frac{f_1(x)}{f_0(x)}$. In particular, when $(f_0, f_1) = (f_{0L}, f_{1L})$ and $g^* = g_R = \log \frac{f_{L1}(x)}{f_{L0}(x)}$, we have the first equality in (3.4). From the second equality, we see that a saddlepoint solution $(g_R, (f_{0L}, f_{1L}))$ exists when (f_{0L}, f_{1L}) is the pair which minimizes $\eta(g_R; f_0, f_1)$ over all $(f_0, f_1) \in \mathcal{F}_0 \times \mathcal{F}_1$; that is, when (f_{0L}, f_{1L}) is the least favorable. Let this be so, and the proof is complete. ∎

In [2], Broder showed that when g is the log-likelihood, then $\eta = \tilde{\eta} = I(f_1, f_0)$, where

$$I(f_1, f_0) = \int_{-\infty}^{\infty} \log\left(\frac{f_1(x)}{f_0(x)}\right) f_1(x) dx$$

is the Kullback-Leibler (K-L) divergence. Therefore,

$$\min_{(f_0, f_1) \in \mathcal{F}_0 \times \mathcal{F}_1} \eta(g_R; f_0, f_1) = \min_{(f_0, f_1) \in \mathcal{F}_0 \times \mathcal{F}_1} I(f_1, f_0)$$

and so the least favorable distribution pair is that which minimizes the K-L divergence.

The goal now is to determine the pair that minimizes $I(f_1, f_0)$. The following Lemma is useful for this purpose. It is almost identical to one which appears in [1].

Lemma 1: *Suppose $\mathcal{P}_0$ and $\mathcal{P}_1$ are classes of probability density functions such that all members of $\mathcal{P}_0 \bigcup \mathcal{P}_1$ have the same support; if $q_0 \in \mathcal{P}_0$ and $q_1 \in \mathcal{P}_1$ are the least favorable in terms of risk for $\mathcal{P}_0$ versus $\mathcal{P}_1$, then*

$$I(q_1, q_0) \leq I(p_1, p_0) \quad \forall p_1 \in \mathcal{P}_1, \forall p_0 \in \mathcal{P}_0$$

The proof is similar to Theorem 2.2 in [1], except that it is slightly more direct and uses the current notation. It will be useful to use the following theorem which appears in [9] and rephrases part of Theorem 2.2 in [1]:

Theorem 1: *Suppose* (p_0, p_1) *and* (q_0, q_1) *are two pairs of probability density functions all of which have the same support. Then, with* $b_1 = 1 - b_0$, *we have*

$$b_0 \int_{\{b_1 q_1 \geq b_0 q_0\}} q_0(x)dx + b_1 \int_{\{b_1 q_1 < b_0 q_0\}} q_1(x)dx$$

$$\geq \; b_0 \int_{\{b_1 p_1 \geq b_0 p_0\}} p_0(x)dx + b_1 \int_{\{b_1 p_1 < b_0 p_0\}} p_1(x)dx \tag{3.5}$$

for all $b_0 \in [0, 1]$ *if and only if*

$$\int_{-\infty}^{\infty} \psi\left[\frac{q_1(x)}{q_0(x)}\right] q_0(x)dx \geq \int_{-\infty}^{\infty} \psi\left[\frac{p_1(x)}{p_0(x)}\right] p_0(x)dx$$

for all continuous concave functions ψ.

Proof: (Lemma 1)

Let $\pi_0 = 1 - \pi_1$ denote the prior probability that $p \in \mathcal{P}_0$. To prove the Lemma, we simply need to let $b_0 = \pi_0$, $b_1 = \pi_1$, and $\psi(u) = -u \log u$. We then directly have that

$$\pi_0 \int_{\{\pi_1 q_1 \geq \pi_0 q_0\}} q_0(x)dx + \pi_1 \int_{\{\pi_1 q_1 < \pi_0 q_0\}} q_1(x)dx$$

$$\geq \; \pi_0 \int_{\{\pi_1 p_1 \geq \pi_0 p_0\}} p_0(x)dx + \pi_1 \int_{\{\pi_1 p_1 < \pi_0 p_0\}} p_1(x)dx$$

which is exactly the condition for $q_0 \in \mathcal{P}_0$ and $q_1 \in \mathcal{P}_1$ to be the least favorable densities in terms of risk for deciding between $\mathcal{P}_0$ and $\mathcal{P}_1$. Using Theorem 1, this implies

$$-\int_{-\infty}^{\infty} \frac{q_1(x)}{q_0(x)} \log\left(\frac{q_1(x)}{q_0(x)}\right) q_0(x)dx \geq -\int_{-\infty}^{\infty} \frac{p_1(x)}{p_0(x)} \log\left(\frac{p_1(x)}{p_0(x)}\right) p_0(x)dx$$

which is the same as $I(q_1, q_0) \leq I(p_1, p_0)$. ∎

The above Lemma establishes a direct relationship between the hypothesis testing problem and the quickest detection problem. In particular, *if a pair of densities is least favorable in terms of Bayes risk, then they will also be the least favorable for the quickest detection problem.* The least favorable pair in terms of risk has been derived for several uncertainty classes [5]; therefore, previous results on robust hypothesis testing can be applied directly to the quickest detection problem.

3.3 Robust Quickest Detectors for Two Noise Uncertainty Models

In this section, the robust quickest detector is determined for two noise uncertainty classes: the *ε-contamination* class, which is a useful noise model when outliers are present in the data, and the *total variation* class, which assumes the noise distribution is of some nominal shape plus or minus some deviation whose sum total does not exceed some constant. The choice of noise model is dependent on the particular application.

For each noise model, we are interested in comparing the performance of the procedures arising from three different design philosophies. The first is the procedure which is optimal for the nominal distributions; if this test works well over the entire class, then a robust approach may not be necessary. Thus, how the test performs when the noise is not generated by the nominal distribution is of primary interest. The second is the minimax robust procedure. The minimax criteria maximizes the performance for the least favorable distributions; however, what sacrifice is made if the noise arises from the nominal distribution? This question is also answered. Finally, the performances for the nonparametric procedures with g chosen to be the sign detector and dead-zone limiter are computed; this enables us to determine when a nonparametric test gives satisfactory performance, such that the use of a robust

procedure is not necessary.

3.3.1 ε-contaminated Noise Class

The noise classes for the ε-contaminated case are as follows:

$$\begin{aligned}\mathcal{F}_0 &= \{f(x) = (1-\varepsilon_0)f_{n0}(x) + \varepsilon_0 h(x) \quad \forall x \in \Re, h \in \mathcal{H}\} \\ \mathcal{F}_1 &= \{f(x) = (1-\varepsilon_1)f_{n1}(x) + \varepsilon_1 h(x) \quad \forall x \in \Re, h \in \mathcal{H}\}\end{aligned}$$

where $\mathcal{H}$ is the class of all legitimate density functions on $\Re$, and ε_0 and ε_1 are constants that lie in the interval $(0,1)$. f_{n0} and f_{n1} are nominal densities in each class. A popular member of the ε-contaminated class is the Gauss-Gauss mixture density:

$$f_{gg}(x) = (1-\varepsilon)\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{\frac{-x^2}{2\sigma_0^2}\right\} + \varepsilon\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{\frac{-x^2}{2\sigma_1^2}\right\}$$

Here, most of the noise samples are Gaussian with variance σ_0^2. However, the noise is sometimes (with probability ε) contaminated by samples which are Gaussian with variance $\sigma_1^2 \gg \sigma_0^2$. This density is useful for modelling noise that is occasionally contaminated by outliers: σ_1^2 and ε are directly proportional to the magnitude and frequency of the outliers, respectively.

Huber [5] has shown that the least favorable densities in terms of risk for the ε-contaminated class are:

$$\begin{aligned}f_{L0}(x) &= \begin{cases}(1-\varepsilon_0)f_{n0}(x), & \frac{f_{n1}(x)}{f_{n0}(x)} < c_0 \\ c_0^{-1}(1-\varepsilon_0)f_{n1}(x), & \frac{f_{n1}(x)}{f_{n0}(x)} \geq c_0\end{cases} \\ f_{L1}(x) &= \begin{cases}(1-\varepsilon_1)f_{n1}(x), & \frac{f_{n1}(x)}{f_{n0}(x)} > c_1 \\ c_1(1-\varepsilon_1)f_{n0}(x), & \frac{f_{n1}(x)}{f_{n0}(x)} \leq c_1\end{cases}\end{aligned}$$

where c_0 and c_1 satisfy $0 < c_1 \leq 1 \leq c_0 < \infty$ and are selected so that f_{L0} and f_{L1} are legitimate density functions; in other words, c_0 and c_1 satisfy

$$\Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} < c_0 \mid f_{n0}\right\} + c_0^{-1}\Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} \geq c_0 \mid f_{n1}\right\} = \frac{1}{1-\varepsilon_0} \qquad (3.6)$$

$$\Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} > c_1 \mid f_{n1}\right\} + c_1 \Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} \leq c_1 \mid f_{n0}\right\} = \frac{1}{1-\varepsilon_1} \tag{3.7}$$

It is shown in [5] that c_0 and c_1 are unique. Furthermore, the left hand sides of (3.6) and (3.7) are both monotonic functions of c_0 and c_1, respectively, so they can be solved using a bisection algorithm. It can be seen directly that the robust nonlinearity for this case is

$$g_R(x) = \log \frac{f_{L1}(x)}{f_{L0}(x)} = \begin{cases} \log c_1 + \log\left(\frac{1-\varepsilon_1}{1-\varepsilon_0}\right), & \frac{f_{n1}(x)}{f_{n0}(x)} \leq c_1 \\ \log \frac{f_{n1}(x)}{f_{n0}(x)} + \log\left(\frac{1-\varepsilon_1}{1-\varepsilon_0}\right), & c_1 < \frac{f_{n1}(x)}{f_{n0}(x)} < c_0 \\ \log c_0 + \log\left(\frac{1-\varepsilon_1}{1-\varepsilon_0}\right), & \frac{f_{n1}(x)}{f_{n0}(x)} \geq c_0 \end{cases}$$

which is seen to be a "censored" version of $\log \frac{f_{n1}(x)}{f_{n0}(x)}$.

Optimal Performance for the Least Favorable Distributions

As shown in the previous section:

$$\eta(g_R; f_{0L}, f_{1L}) = I(f_{L1}, f_{L0})$$

Thus, the asymptotic performance is obtained by simply evaluating the K-L divergence for the least favorable distributions. The result is

$$\begin{aligned}
\eta(g_R; f_{0L}, f_{1L}) &= \int_{\left\{\frac{f_{n1}(x)}{f_{n0}(x)} \leq c_1\right\}} \left[\log c_1 + \log\left(\frac{1-\varepsilon_1}{1-\varepsilon_0}\right)\right] c_1(1-\varepsilon_1) f_{n0}(x)dx \\
&+ \int_{\left\{c_1 < \frac{f_{n1}(x)}{f_{n0}(x)} < c_0\right\}} \left[\log\left(\frac{f_{n1}(x)}{f_{n0}(x)}\right) + \log\left(\frac{1-\varepsilon_1}{1-\varepsilon_0}\right)\right] (1-\varepsilon_1) f_{n1}(x)dx \\
&+ \int_{\left\{\frac{f_{n1}(x)}{f_0(x)} \geq c_0\right\}} \left[\log c_0 + \log\left(\frac{1-\varepsilon_1}{1-\varepsilon_0}\right)\right] (1-\varepsilon_1) f_{n1}(x)dx \\
&= (1-\varepsilon_1)\log\left(\frac{1-\varepsilon_1}{1-\varepsilon_0}\right)\left[c_1 \Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} \leq c_1 \mid f_{n0}\right\} + \Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} > c_1 \mid f_{n1}\right\}\right] \\
&+ (1-\varepsilon_1)\left[c_1 \log c_1 \Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} \leq c_1 \mid f_{n0}\right\} + \log c_0 \Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} \geq c_0 \mid f_{n1}\right\}\right] \\
&+ (1-\varepsilon_1) \int_{\left\{c_1 < \frac{f_{n1}(x)}{f_{n0}(x)} < c_0\right\}} \log\left(\frac{f_{n1}(x)}{f_{n0}(x)}\right) f_{n1}(x)dx
\end{aligned}$$

$$= \log\left(\frac{1-\varepsilon_1}{1-\varepsilon_0}\right) + (1-\varepsilon_1)\int_{\left\{c_1 < \frac{f_{n1}(x)}{f_{n0}(x)} < c_0\right\}} \log\left(\frac{f_{n1}(x)}{f_{n0}(x)}\right) f_{n1}(x)dx$$

$$+ (1-\varepsilon_1)\left[c_1 \log c_1 \Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} \leq c_1 \mid f_{n0}\right\} + \log c_0 \Pr\left\{\frac{f_{n1}(x)}{f_{n0}(x)} \geq c_0 \mid f_{n1}\right\}\right] \quad (3.8)$$

where the last inequality results by substituting in (3.7). The ε-contamination model shown above makes the provision that the contamination of the nominal distributions may differ under the null and alternative hypotheses, and so $\varepsilon_0 \neq \varepsilon_1$. While there are some applications where this is the case, the present work focuses on the case the disorder is solely due to a shift in the mean. Hence, we let $\varepsilon_0 = \varepsilon_1 = \varepsilon$.

ε-contaminated Model with Nominal Gaussian Noise: Optimal Performance for the Least Favorable Distributions

Here, we consider the specific case where the noise pair lies in the ε-contaminated class with nominal Gaussian distributions, and $\varepsilon_0 = \varepsilon_1 = \varepsilon$. Let

$$\varphi(x;\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$$

and define $\phi_0(x) = \varphi(x+\theta;\sigma)$ and $\phi_1(x) = \varphi(x-\theta;\sigma)$. The nominal densities are now $f_{n0}(x) = \phi_0(x)$ and $f_{n1}(x) = \phi_1(x)$.

In Figure 3.1, examples of the least favorable pair are shown for $\varepsilon = 0.1$ and $\varepsilon = 0.25$ along with the nominal pair, where $\theta = \sigma_0^2 = 1$. Notice that the least favorable contamination of f_{n0} involves increasing the density in the neighborhood of f_{n1}, and vice versa, making it more difficult to distinguish between the two hypotheses.

The optimal performance for the least favorable distribution pair (f_{n0}, f_{n1}) is given in (3.8). Compute the four terms individually. The first one is clearly zero. Computing the second term requires us to evaluate the integral

$$\int_{\left\{c_1 < \frac{\phi_1(x)}{\phi_0(x)} < c_0\right\}} \log\left(\frac{\phi_1(x)}{\phi_0(x)}\right) \phi_1(x)dx = \frac{2\theta}{\sigma_0^2} \int_{\frac{\sigma_0^2}{2\theta}\log c_1 - \theta}^{\frac{\sigma_0^2}{2\theta}\log c_0 - \theta} \frac{y+\theta}{\sqrt{2\pi}\sigma_0} \exp\left\{\frac{-y^2}{2\sigma_0^2}\right\} dy \quad (3.9)$$

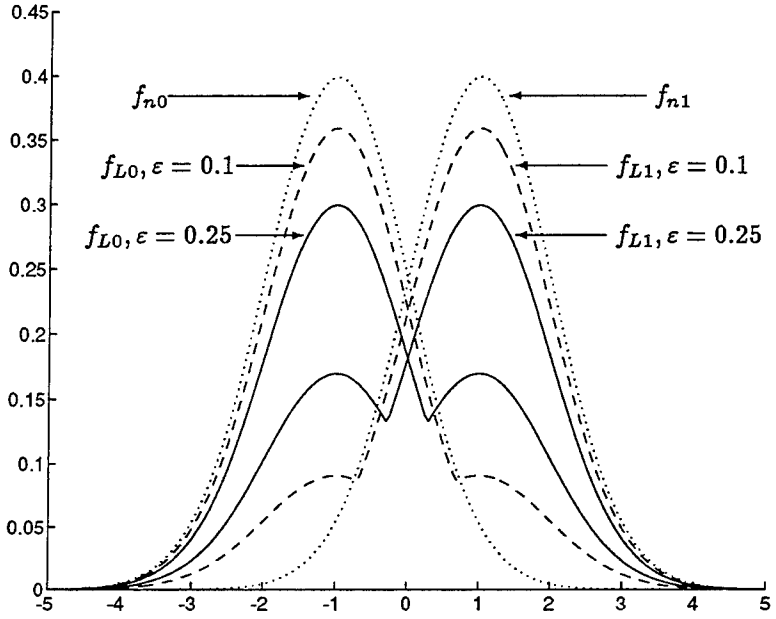

Figure 3.1: Nominal Gaussian and least favorable ε-contaminated densities. $\theta = \sigma_0^2 = 1$.

where the change of variables $y = x - \theta$ was made. The right side can be broken up into two parts:

$$\frac{2\theta}{\sigma_0^2}\int_{b_1-\theta}^{b_0-\theta}\frac{y}{\sqrt{2\pi}\sigma_0}\exp\left\{\frac{-y^2}{2\sigma_0^2}\right\}dy + \frac{2\theta^2}{\sigma_0^2}\int_{b_1-\theta}^{b_0-\theta}\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{\frac{-y^2}{2\sigma_0^2}\right\}dy$$

where $b_0 \triangleq \frac{\sigma_0^2}{2\theta}\log c_0$ and $b_1 \triangleq \frac{\sigma_0^2}{2\theta}\log c_1$. The left term can be integrated directly and the right term can be written in terms of the standard normal cumulative distribution function Φ. Therefore, (3.9) equals

$$\sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma_0}\left[\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\} - \exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}\right] + \frac{2\theta^2}{\sigma_0^2}\left[\Phi\left(\frac{b_0-\theta}{\sigma_0}\right) - \Phi\left(\frac{b_1-\theta}{\sigma_0}\right)\right]$$

Evaluating the third term requires the following computation:

$$\Pr\left\{\frac{\phi_1(x)}{\phi_0(x)} \le c_1 \mid \phi_0\right\} = \Pr\left\{x \le \frac{\sigma_0^2}{2\theta}\log c_1 \mid \phi_0\right\} = \Phi\left(\frac{b_1+\theta}{\sigma_0}\right) \tag{3.10}$$

Similarly, to evaluate the fourth term, we need:

$$\Pr\left\{\frac{\phi_1(x)}{\phi_0(x)} \ge c_0 \mid \phi_1\right\} = \Pr\left\{x \ge \frac{\sigma_0^2}{2\theta}\log c_0 \mid \phi_1\right\} = 1 - \Phi\left(\frac{b_0-\theta}{\sigma_0}\right) = \Phi\left(\frac{-b_0+\theta}{\sigma_0}\right) \tag{3.11}$$

By using (3.9)-(3.11), we finally have that (3.8) is

$$\begin{aligned}\eta(g_R; f_{0L}, f_{1L}) &= (1-\varepsilon)\left\{\sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma_0}\left[\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\} - \exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}\right]\right\} \\ &+ (1-\varepsilon)\left\{\frac{2\theta^2}{\sigma_0^2}\left[\Phi\left(\frac{b_0-\theta}{\sigma_0}\right) - \Phi\left(\frac{b_1-\theta}{\sigma_0}\right)\right]\right\} \\ &+ (1-\varepsilon)\left\{c_1 \log c_1\, \Phi\left(\frac{b_1+\theta}{\sigma_0}\right) + \log c_0\, \Phi\left(\frac{-b_0+\theta}{\sigma_0}\right)\right\}\end{aligned}$$

where c_0 and c_1 are chosen to satisfy

$$\Phi\left(\frac{b_0+\theta}{\sigma_0}\right) + c_0^{-1}\Phi\left(\frac{-b_0+\theta}{\sigma_0}\right) = \frac{1}{1-\varepsilon} \tag{3.12}$$

$$\Phi\left(\frac{-b_1+\theta}{\sigma_0}\right) + c_1\Phi\left(\frac{b_1+\theta}{\sigma_0}\right) = \frac{1}{1-\varepsilon} \tag{3.13}$$

Three interesting points can be made about the above equations. First, notice that $\eta(g_R; f_{0L}, f_{1L})$ can be alternatively written in terms of the nominal signal to noise ratio $d \triangleq \frac{\theta}{\sigma_0}$:

$$\begin{aligned}\eta &= (1-\varepsilon)\left\{\sqrt{\frac{2}{\pi}}d\left[\exp\left\{-\frac{1}{2}\left(\frac{1}{2d}\log c_1 - d\right)^2\right\} - \exp\left\{-\frac{1}{2}\left(\frac{1}{2d}\log c_0 - d\right)^2\right\}\right]\right\} \\ &+ (1-\varepsilon)\left\{2d^2\left[\Phi\left(\frac{1}{2d}\log c_0 - d\right) - \Phi\left(\frac{1}{2d}\log c_1 - d\right)\right]\right\} \\ &+ (1-\varepsilon)\left\{c_1 \log c_1\, \Phi\left(\frac{1}{2d}\log c_1 + d\right) + \log c_0\, \Phi\left(-\frac{1}{2d}\log c_0 + d\right)\right\}\end{aligned}$$

This demonstrates that the performance is proportional to the *relative* values of the nominal signal θ and nominal noise power σ_0^2, rather than the *absolute* values. This is not surprising because: *i)* we are using Gaussian nominals, and *ii)* for hypothesis testing in Gaussian noise, the probability of error is a function of the signal to noise ratio [11].

Another observation is that $c_0 = c_1^{-1}$ (and, hence, $b_0 = -b_1$). To see this, first assume that c_0 satisfies (3.12), and then substitute $c_0 \leftarrow c_1^{-1}$; the result is equation (3.13). Similar logic shows that (3.12) can be obtained from (3.13). Since both equations must hold, $c_0 = c_1^{-1}$ (note that this is only the case when $\varepsilon_0 = \varepsilon_1$).

The last point involves the range of values that ε can take on. An upper bound can be obtained by considering the case when $c_0 = c_1$, corresponding to the minimum value of c_0 for which the model is still valid (recall that $c_1 \leq c_0$). Since $c_0 \geq 0$, this occurs at $c_0 = c_1 = 1$. [5] Substituting these values into (3.12) and (3.13), we have $\varepsilon = \varepsilon^*$, where

$$\varepsilon^* = 1 - \frac{1}{2}\left[\Phi\left(\frac{\theta}{\sigma_0}\right)\right]^{-1}$$

Also notice that (3.13) can be written as

$$\Phi\left(\frac{-b_1 + \theta}{\sigma_0}\right) + \exp\left\{\frac{2\theta b_1}{\sigma_0^2}\right\}\Phi\left(\frac{b_1 + \theta}{\sigma_0}\right) = \frac{1}{1-\varepsilon}$$

and that the left side of this equation is monotonically increasing in $b_1 = -b_0 < 0$. Therefore, the largest allowable value of b_1, namely zero, results in the largest allowable value of ε, namely ε^*. We have thus defined the *breakdown point*, ε^*, for the ε-contaminated noise class: the condition $\varepsilon \in [0, \varepsilon^*)$ must be met for the problem to be valid. In other words, ε is small enough to insure that $\mathcal{F}_0$ and $\mathcal{F}_1$ are distinguishable. The breakdown point is plotted versus $\frac{\theta}{\sigma_0}$ in Figure 3.2.

Performance Involving the ε-contaminated Model with Nominal Gaussian Noise when the Assumed and True Noise Densities Differ

Perhaps the most interesting performance computations involve cases where the noise assumptions used to design the detector do not match the true distributions. Examining such cases enables us to evaluate the performance degradation that results when it is not possible to implement the optimal procedure, for example, when the noise is not completely characterized. It was shown in [2] that when the assumed and true distributions, say f_0 and f_1, agree, then $\eta = \tilde{\eta} = I(f_1, f_0)$. However, in general, not only is the computation of η is more complicated, but also the lower bound, $\tilde{\eta}$, on η is not tight. We will see, though, that $\tilde{\eta}$ is still a good approximation for η.

[5] Note, however, that this results in the degenerate case of $g_R(x) \equiv 0$.

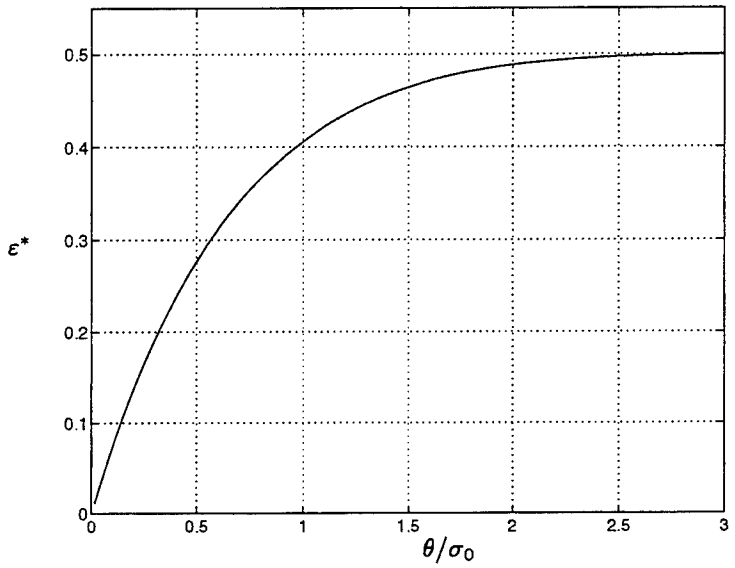

Figure 3.2: Breakdown point for the ε-contaminated noise model.

The procedure for computing $\tilde{\eta}$, detailed in Chapter 2, is reiterated here. Let $g(x)$ be an arbitrary nonlinearity, and let (f_0, f_1) denote the true noise density pair. Then $\tilde{\eta}$ is given by

$$\tilde{\eta} = \omega_0 \mathrm{E}\{g(x) \mid f_1\} \tag{3.14}$$

where ω_0 is the unique nonzero root of the moment generating function equality

$$\mathrm{E}\left\{e^{w_0 g(x)} \mid f_0\right\} = 1 \tag{3.15}$$

When $g(x)$ is the log-likelihood ratio between f_0 and f_1, $\omega_0 = 1$ and (3.14) reduces to the K-L divergence. If this is not the case, (3.15) must be solved for ω_0, and then (3.14) can be computed.

Table 1 summarizes each of the scenarios for which $\tilde{\eta}$ is computed. For the linear and robust detectors, the particular expressions corresponding to equations (3.14) and (3.15) can be found in Appendix A. The computation of $\tilde{\eta}$ for nonparametric procedures is discussed next.

Table 1: Computations involving the ε-contaminated model

noise type	*procedure used*
Gaussian	linear
Gauss-Gauss	sign
least-favorable	dead-zone
	robust

Nonparametric Alternatives

The robust quickest detector is a *parametric* procedure. That is, specific assumptions are made about the noise classes: they are "centered" about some known nominal distribution, and the contamination factor ε is known. On the other hand, *nonparametric* detectors involve procedures that make no assumptions on the noise characteristics, but assume only that a shift in the mean from some negative value to some positive value occurs. It is useful to compare the performance of the quickest detectors arising from these two paradigms. Nonparametric quickest detection has been considered extensively in [2]; the main results of that work are repeated here.

Suppose the nonlinearity is given by

$$g(x) = \begin{cases} -1, & x < -d \\ 0, & -d \le x \le d \\ 1, & x > d \end{cases}$$

This is called a *random walk nonlinearity.* In [2] it is shown that when g is used in Page's test, the lower bound $\tilde{\eta}$ is tight, and is given by

$$\tilde{\eta} = [p_1 - q_1] \log \frac{q_0}{p_0}$$

where

$$\begin{aligned} p_i &= \Pr\{g(x) = 1 \mid f_i\}, \quad i = 0, 1 \\ q_i &= \Pr\{g(x) = -1 \mid f_i\}, \quad i = 0, 1 \end{aligned}$$

When $d = 0$, $g(x)$ is the sign detector; otherwise, it is called a dead-zone limiter. For the latter case, a value of $d = 0.612 \cdot \sigma_0$ is used, since this choice maximizes the efficacy given that the noise is Gaussian [6]; this is reasonable since the noise is nominally Gaussian with variance σ_0^2. [6] Once f_0 and f_1 are known, the values p_i and q_i are easily computed, and hence, so is the performance.

We would like to determine what level of robustness can be gained by using nonparametric techniques. It was shown in [2] that the sign detector and dead-zone limiter often outperform the linear detector when the tails of the noise distribution are heavier than Gaussian. Intuitively, one expects that the robust procedures would outperform their nonparametric counterparts, since more assumptions are incorporated into the model: as we shall see, this is true in most cases.

3.3.2 Total Variation Noise Class

As in Section 3.3.1, denote the nominal density function pair by (f_{n0}, f_{n1}) and the least favorable pair by (f_{L0}, f_{L1}). The noise classes for the total variation model are:

$$\begin{aligned} \mathcal{F}_0 &= \{f : \int_{-\infty}^{\infty} |f(x) - f_{n0}(x)|\, dx \leq \delta\} \\ \mathcal{F}_1 &= \{f : \int_{-\infty}^{\infty} |f(x) - f_{n1}(x)|\, dx \leq \delta\} \end{aligned}$$

That is, the sum total of the variation of $f_i \in \mathcal{F}_i$ from the nominal f_{ni} does not exceed δ. This class is useful in cases where the overall shape of the noise density is exactly or approximately known, but where there is still some uncertainty proportional to δ. This uncertainty might arise due to modelling error, or possibly due to assuming that the noise is stationary when in fact it is nonstationary.

[6] In [2], the dead-zone limiter is implemented with $d = 0.612 \cdot \sigma$, where σ^2 is the variance of the f_i rather than f_{ni}. It makes no sense to do this here, however, since we do not know what the noise density is in reality, only that the nominal density is $\mathcal{N}(0, \sigma_0^2)$.

The least favorable densities in this case are [5]:

$$f_{L0}(x) = \begin{cases} \frac{1}{1+c_0}(f_{n0}(x)+f_{n1}(x)), & \frac{f_{n1}(x)}{f_{n0}(x)} \le c_0 \\ f_{n0}(x), & c_0 < \frac{f_{n1}(x)}{f_{n0}(x)} < c_1 \\ \frac{1}{1+c_1}(f_{n0}(x)+f_{n1}(x)), & \frac{f_{n1}(x)}{f_{n0}(x)} \ge c_1 \end{cases}$$

$$f_{L1}(x) = \begin{cases} \frac{c_0}{1+c_0}(f_{n0}(x)+f_{n1}(x)), & \frac{f_{n1}(x)}{f_{n0}(x)} \le c_0 \\ f_{n1}(x), & c_0 < \frac{f_{n1}(x)}{f_{n0}(x)} < c_1 \\ \frac{c_1}{1+c_1}(f_{n0}(x)+f_{n1}(x)), & \frac{f_{n1}(x)}{f_{n0}(x)} \ge c_1 \end{cases}$$

and the robust nonlinearity is

$$g_R(x) = \log \frac{f_{L1}(x)}{f_{L0}(x)} = \begin{cases} \log c_0, & \frac{f_{n1}(x)}{f_{n0}(x)} \le c_0 \\ \log \frac{f_{n1}(x)}{f_{n0}(x)}, & c_0 < \frac{f_{n1}(x)}{f_{n0}(x)} < c_1 \\ \log c_1, & \frac{f_{n1}(x)}{f_{n0}(x)} \ge c_1 \end{cases}$$

where $0 < c_0 \le 1 \le c_1 < \infty$, which again is seen to be a "censored" version of $\log \frac{f_{n1}(x)}{f_{n0}(x)}$. [7]

The least favorable distributions satisfy

$$\int_{-\infty}^{\infty} |f_{L0}(x) - f_{n0}(x)|\, dx = \delta$$
$$\int_{-\infty}^{\infty} |f_{L1}(x) - f_{n1}(x)|\, dx = \delta$$

It is shown in [5] that a sufficient condition for this is that

$$\int_{\left\{\frac{f_{n1}}{f_{n0}} \le c_0\right\}} (f_{L1}(x) - f_{n1}(x))dx = \int_{\left\{\frac{f_{n1}}{f_{n0}} \ge c_1\right\}} (f_{L0}(x) - f_{n0}(x))dx = \frac{\delta}{2} \qquad (3.16)$$

The values of c_0 and c_1 are determined by solving (3.16). As with the ε-contamination noise model, c_0 and c_1 are unique. In particular, if we let $k_0 = \frac{c_0}{1+c_0}$, then the first term in (3.16) can be written as

$$\int_{\{f_{n1} \le (f_{n0}+f_{n1})k_0\}} [(f_{n0}(x)+f_{n1}(x))k_0 - f_{n1}(x)]\, dx$$

[7] Note that here $c_0 \le c_1$, whereas in Section 3.1, $c_1 \le c_0$.

which is an increasing function of k_0; thus, a bisection algorithm can be used to solve for k_0, and hence, for c_0 as well. Similarly, c_1 is determined by letting $k_1 = \frac{1}{1+c_1}$ and solving

$$\int_{\{f_{n0} \le (f_{n0}+f_{n1})k_1\}} [(f_{n0}(x) + f_{n1}(x))k_1 - f_{n0}(x)]\, dx$$

which is an increasing function of k_1. The above conditions can be written in the more useful form:

$$k_0 \Pr\left\{\frac{f_{n1}}{f_{n0}} \le c_0 \mid f_{n0}\right\} + (k_0 - 1)\Pr\left\{\frac{f_{n1}}{f_{n0}} \le c_0 \mid f_{n1}\right\} = \frac{\delta}{2} \qquad (3.17)$$

$$(k_1 - 1)\Pr\left\{\frac{f_{n1}}{f_{n0}} \ge c_1 \mid f_{n0}\right\} + k_1 \Pr\left\{\frac{f_{n1}}{f_{n0}} \ge c_1 \mid f_{n1}\right\} = \frac{\delta}{2} \qquad (3.18)$$

Optimal Performance for the Least Favorable Distributions

As shown in Section 2, the optimal robust asymptotic performance is given by the K-L divergence, which for the least favorable total variation density pair is

$$\begin{aligned}
\eta(g_R; f_{0L}, f_{1L}) &= \int_{\left\{\frac{f_{n1}}{f_{n0}} \le c_0\right\}} (\log c_0)\frac{c_0}{1+c_0}(f_{n0}(x) + f_{n1}(x))dx \\
&+ \int_{\left\{c_0 < \frac{f_1}{f_0} < c_1\right\}} \log\left(\frac{f_{n1}(x)}{f_{n0}(x)}\right) f_{n1}(x)dx \\
&+ \int_{\left\{\frac{f_{n1}}{f_{n0}} \ge c_1\right\}} (\log c_1)\frac{c_1}{1+c_1}(f_{n0}(x) + f_{n1}(x))dx \\
&= \frac{c_0}{1+c_0}\log c_0 \left[\Pr\left\{\frac{f_{n1}}{f_{n0}} \le c_0 \mid f_{n0}\right\} + \Pr\left\{\frac{f_{n1}}{f_{n0}} \le c_0 \mid f_{n1}\right\}\right] \\
&+ \frac{c_1}{1+c_1}\log c_1 \left[\Pr\left\{\frac{f_{n1}}{f_{n0}} \ge c_1 \mid f_{n0}\right\} + \Pr\left\{\frac{f_{n1}}{f_{n0}} \ge c_1 \mid f_{n1}\right\}\right] \\
&+ \int_{\left\{c_0 < \frac{f_{n1}}{f_{n0}} < c_1\right\}} \log\left(\frac{f_{n1}(x)}{f_{n0}(x)}\right) f_{n1}(x)dx
\end{aligned}$$

Total Variation Model with Nominal Gaussian Noise: Optimal Performance for the Least Favorable Distributions

Here, we consider the specific case where the noise pair lies in the total variation class with nominal Gaussian distributions $f_{n0}(x) = \phi_0(x)$ and $f_{n1}(x) = \phi_1(x)$. In Figure 3.3, the least favorable densities in the total variation class are shown for $\delta = 0.2$ and $\delta = 0.5$ along with the nominal Gaussian densities, where $\theta = \sigma_0^2 = 1$. As with the least favorable ε-contaminated densities, f_{L0} and f_{L1} for the total variation model are seen to look like one nominal corrupted by the other, thus increasing the difficulty in distinguishing $\mathcal{F}_0$ from $\mathcal{F}_1$.

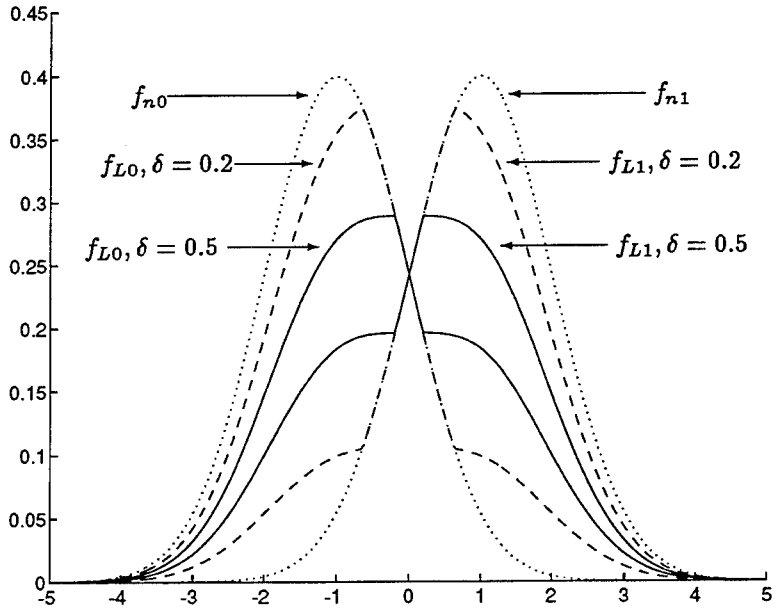

Figure 3.3: Nominal Gaussian and least favorable total variation densities. $\theta = \sigma_0^2 = 1$.

Again, for convenience, define $b_0 \stackrel{\Delta}{=} \frac{\sigma_0^2}{2\theta} \log c_0$ and $b_1 \stackrel{\Delta}{=} \frac{\sigma_0^2}{2\theta} \log c_1$. Then

$$\begin{aligned} \eta(g_R; f_{0L}, f_{1L}) &= \frac{c_0}{1+c_0} \log c_0 \left[\Phi\left(\frac{b_0 + \theta}{\sigma_0} \right) + \Phi\left(\frac{b_0 - \theta}{\sigma_0} \right) \right] \\ &\quad + \frac{c_1}{1+c_1} \log c_1 \left[\Phi\left(\frac{-b_1 - \theta}{\sigma_0} \right) + \Phi\left(\frac{-b_1 + \theta}{\sigma_0} \right) \right] \end{aligned}$$

$$+\sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma_0}\left[\exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}-\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\}\right]$$
$$+\frac{2\theta^2}{\sigma_0^2}\left[\Phi\left(\frac{b_1-\theta}{\sigma_0}\right)-\Phi\left(\frac{b_0-\theta}{\sigma_0}\right)\right]$$

The last two terms arise when the integral over $\{c_0 < \frac{f_1}{f_0} < c_1\}$ is evaluated; however, note that this is the same computation as in (3.9), except that the roles of c_0 and c_1 are now reversed. The values of c_0 and c_1 are determined by solving (3.17) and (3.18) for this particular choice of densities. They are:

$$\begin{aligned} k_0\Phi\left(\frac{b_0+\theta}{\sigma_0}\right)+(k_0-1)\Phi\left(\frac{b_0-\theta}{\sigma_0}\right) &= \frac{\delta}{2} \\ (k_1-1)\Phi\left(\frac{-b_1-\theta}{\sigma_0}\right)+k_1\Phi\left(\frac{-b_1+\theta}{\sigma_0}\right) &= \frac{\delta}{2} \end{aligned}$$

(Recall that b_i and k_i are functions of c_i, for $i = 0, 1$)

As with the ε-contaminated model, $\eta(g_R; f_{0L}, f_{1L})$ for the total variation model can be expressed in terms of the nominal signal to noise ratio $d \triangleq \frac{\theta}{\sigma_0}$. We have

$$\begin{aligned} \eta(g_R; f_{0L}, f_{1L}) &= \frac{c_0}{1+c_0}\log c_0\left[\Phi\left(\frac{1}{2d}\log c_0+d\right)+\Phi\left(\frac{1}{2d}\log c_0-d\right)\right] \\ &+ \frac{c_1}{1+c_1}\log c_1\left[\Phi\left(-\frac{1}{2d}\log c_1-d\right)+\Phi\left(-\frac{1}{2d}\log c_1+d\right)\right] \\ &+ \sqrt{\frac{2}{\pi}}d\left[\exp\left\{-\frac{1}{2}\left(\frac{1}{2d}\log c_0-d\right)^2\right\}-\exp\left\{-\frac{1}{2}\left(\frac{1}{2d}\log c_1-d\right)^2\right\}\right] \\ &+ 2d^2\left[\Phi\left(\frac{1}{2d}\log c_1-d\right)-\Phi\left(\frac{1}{2d}\log c_0-d\right)\right] \end{aligned}$$

Also, since the nominal densities are symmetric, we again have that $c_0 = c_1^{-1}$.

The breakdown point can be computed in a manner similar to before. We notice that again this occurs at $c_0 = c_1 = 1$, and so the critical points for δ are

$$\delta^* = \Phi\left(\frac{\theta}{\sigma_0}\right)-\Phi\left(\frac{-\theta}{\sigma_0}\right) = 2\Phi\left(\frac{\theta}{\sigma_0}\right)-1$$

Hence, for a given $\frac{\theta}{\sigma_0}$, the model is valid for $\delta \in [0, \delta^*)$. The breakdown point is plotted versus $\frac{\theta}{\sigma_0}$ in Figure 3.4.

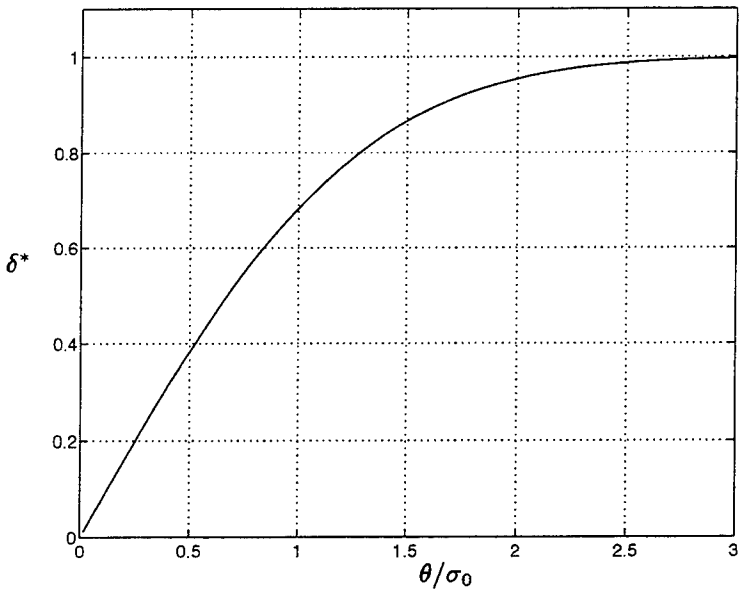

Figure 3.4: Breakdown point for the total variation noise model.

Performance Involving the Total Variation Model with Nominal Gaussian Noise When Assumed and True Noise Densities Differ

Here, $\widetilde{\eta}$ is computed for several cases where the noise assumptions used to design the detector do not match the true distributions. A summary of all computations involving the total variation model is shown in Table 2. The discussion in Section 3.3.1 regarding the computation of $\widetilde{\eta}$ also pertains here. The details of the derivations can be found in Appendix B.

Table 2: Computations involving the total variation model

noise type	*procedures used*
Gaussian	linear
least-favorable	sign
	dead-zone
	robust

Nonparametric Alternatives

The discussion of Section 3.3.1 also applies here. The versions of Page's test involving the sign detector and dead-zone limiter are again evaluated against the robust quickest detector.

3.4 Performance Comparison

In this section, the performances for all of the noise/quickest detector combinations are computed for some particular cases involving the ε-contamination and total variation models. Throughout, we assume for both noise classes that the nominal distribution is Gaussian with variance $\sigma_0^2 = 1$, and that the mean is $-\theta$ before the disorder and θ afterwards.

For each of the two noise classes, the performance will be computed for the cases where the true noise pair (f_0, f_1) is: *i)* the nominal, and *ii)* the least favorable pair. In addition, the performance is also computed for another member of the ε-contaminated class, the Gauss-Gauss mixture, where the contaminating distribution is Gaussian with variance $\sigma_1^2 = 100$. The parameter $\gamma \stackrel{\Delta}{=} \frac{\sigma_1^2}{\sigma_0^2}$ is often used to represent the magnitude of the outliers relative to that of the nominal noise samples; here $\gamma = 100$.

It will be useful to define the *effective* signal to noise ratio (SNR), Ψ, as follows:

$$\Psi = \frac{|\mathrm{E}\{X \mid f_i\}|}{\sqrt{\mathrm{Var}\{X \mid f_i\}}}, \quad f_i \in \mathcal{F}_i, \quad i = 0, 1$$

This is seen to be a weighting of the signal strength (the mean) by the uncertainty (the variance); both of these quantities vary for different members of $\mathcal{F}_i$. Because we are considering the case where the nominal densities are symmetric, that is, $f_{n0}(-x) = f_{n1}(x)$, we see that $\mathrm{E}\{X \mid f_1\} = -\mathrm{E}\{X \mid f_0\}$. Thus, the effective SNR can now be

related to the *deflection* SNR [3]:

$$\mathrm{SNR}_{def} \triangleq \frac{|\mathrm{E}\{X \mid f_1\} - \mathrm{E}\{X \mid f_0\}|^2}{\mathrm{Var}\{X \mid f_0\}}$$

which is a measure of the relative detectability between the pre- and post-disorder hypotheses, f_0 and f_1; [8] we have:

$$\Psi = \frac{1}{2}\left(\mathrm{SNR}_{def}\right)^{\frac{1}{2}}$$

Note also that the effective SNR can be alternatively expressed in decibels:

$$\Psi_{dB} = 10\log_{10}\Psi^2$$

It is not difficult to verify that for Gaussian noise

$$\Psi = \frac{\theta}{\sigma_0}$$

and for Gauss-Gauss noise

$$\Psi = \frac{\theta}{\sqrt{(1-\varepsilon)\sigma_0^2 + \varepsilon\sigma_1^2}}$$

However, the computation of Ψ for the least favorable distributions is more involved, and therefore can be found in Appendix C.

In the previous section, $\tilde{\eta}$ was computed for several noise and detector combinations. We validate the accuracy of those expressions by comparing the computed values of $\tilde{\eta}$ with estimates of η obtained by measuring the inverse of the asymptotic slope of the plot of D versus $\log T$ (recall that this is exactly the definition of η). In

[8] A more general definition of the deflection SNR is

$$\mathrm{SNR}_{def} \triangleq \frac{|\mathrm{E}\{T(X) \mid f_1\} - \mathrm{E}\{T(X) \mid f_0\}|^2}{\mathrm{Var}\{T(X) \mid f_0\}}$$

which is useful in evaluating the power of a detection procedure where the samples are processed by the nonlinearity $T(\cdot)$. Thus, $T(x) = x$ in the present case. A comparison of other definitions of the SNR is the subject of [3].

Appendix D, performance curves are shown for each of the noise distribution pairs for $\Psi_{dB} = 0$ and -10 dB, followed by a series of tables summarizing the measured and computed values of the asymptotic performance. In all cases, η and $\tilde{\eta}$ agree within 2%. Therefore, based on the computations performed, the approximation $\eta \approx \tilde{\eta}$ appears to be well-founded.

In order to evaluate the performance gain or loss in opting to use a robust procedure rather than one that is nonparametric or based on the nominals, we define the *robustness index* as

$$\chi_{A,B} \stackrel{\Delta}{=} \frac{\tilde{\eta}_A}{\tilde{\eta}_B}$$

where $\tilde{\eta}_A$ and $\tilde{\eta}_B$ are the values of $\tilde{\eta}$ for procedures labelled A and B, respectively. The robustness index also allows us to relate the relative expected delays for each procedure. We have seen that $\tilde{\eta}$ closely approximates η. As a result, for large T, equation (3.2) in Section 2 becomes

$$D \approx \frac{\log T}{\tilde{\eta}} \tag{3.19}$$

Let D_A and D_B denote the expected delays for procedures A and B, for some T which is the same for both procedures. Using (3.19), we have

$$\frac{D_B}{D_A} \approx \left(\frac{\log T}{\tilde{\eta}_B}\right) \div \left(\frac{\log T}{\tilde{\eta}_A}\right) = \frac{\tilde{\eta}_A}{\tilde{\eta}_B} = \chi_{A,B}$$

Thus, a decrease in the expected delay corresponds to an increase in the performance.

3.4.1 $\tilde{\eta}$ Versus Ψ for Different Detector/Noise Combinations

Figures 3.5 through 3.18 illustrate the asymptotic performance for each of the procedures as a function of Ψ, as well as the improvement index of the robust procedure over each of the others. Figures 3.5 and 3.6 show the performance when the noise

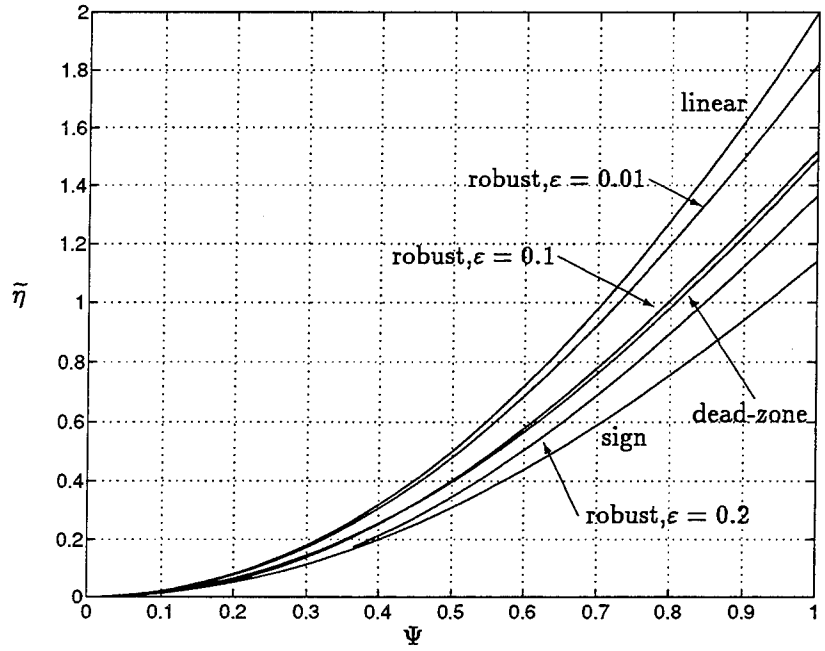

Figure 3.5: $\widetilde{\eta}$ for Gaussian noise.

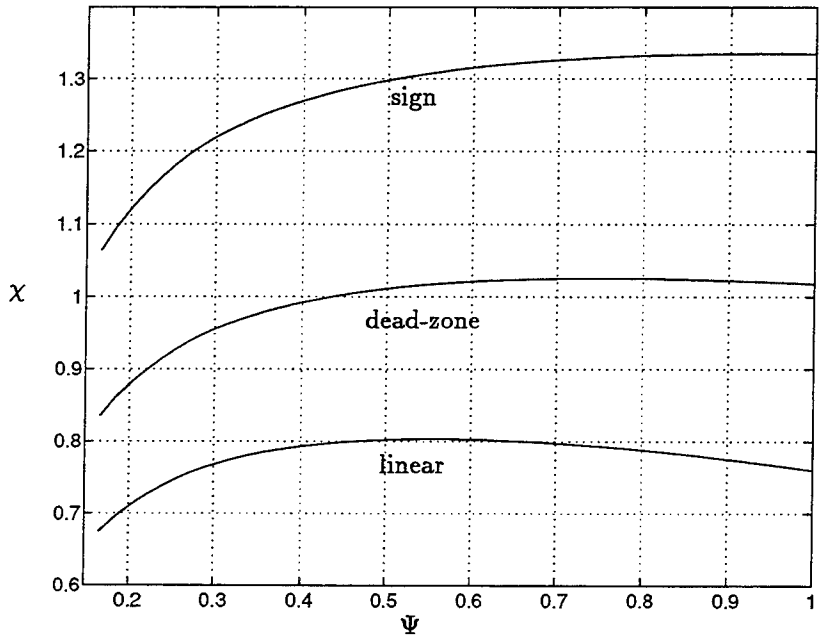

Figure 3.6: Robustness index for the robust procedure with $\varepsilon = 0.1$ in Gaussian noise.

is the nominal Gaussian. [9] The linear detector is optimal for this case, and this is reflected in the plots. Also notice that, when the robust detector is used, the assumed level of contamination ε is inversely proportional to the performance. The plot of χ versus Ψ reveals that the expected delay for the linear detector is only about 75% of that for the robust procedure. This is the price one pays for robustness: when the noise is close to nominal, the robust procedure will react less quickly than the procedure which is optimal for (f_{n0}, f_{n1}). On the other hand, the robust procedure outperforms the sign detector for any choice of ε and the dead-zone limiter for smaller ε; however, notice that if $\varepsilon = 0.2$ is assumed, then the dead-zone limiter is the better choice. We see that as ε gets larger, the robust procedure incorporates less information about the nominals (i.e., the nominal log-likelihood ratio is clipped at lower levels), resulting in a decrease in $\tilde{\eta}$.

On the flip side of the above discussion, the conservativism of the robust procedure can be more than offset when the noise is not nominal, as shown in Figures 3.7 through 3.10. Here, Gauss-Gauss noise is considered for $\varepsilon = 0.01$ and $\varepsilon = 0.1$. The most striking observation is that the linear detector performs much more poorly than any of the other procedures considered: the robust procedure outperforms the linear test by more than a factor of eight in some cases, and would therefore be preferred in a noise environment in which outliers are present. The performances of the nonparametric tests are more reasonable. Also, the relative benefit of the robust over the nonparametric tests in general is smaller for $\varepsilon = 0.1$; in fact, the procedure which utilizes the dead-zone limiter outperforms the robust test for Ψ close to unity.

The performance for least favorable ε-contaminated noise is shown in Figures 3.11 through 3.14. The quickest detector for the least favorable noise is exactly the robust procedure (by design), and the graphs corroborate this. For the case of small

[9] Here χ denotes the gain in performance that is realized when the robust procedure is used compared to the procedure listed on the graph.

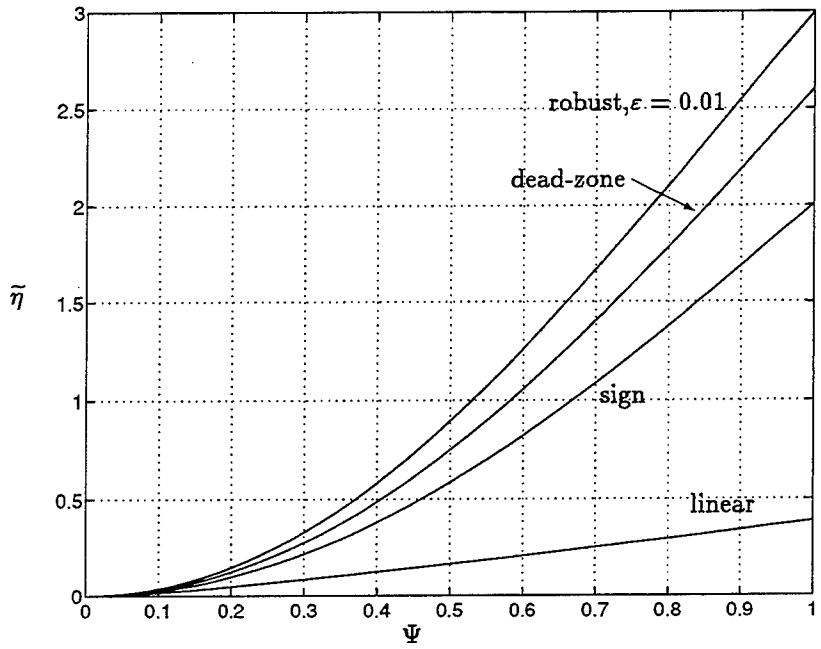

Figure 3.7: $\widetilde{\eta}$ for Gauss-Gauss noise: $\varepsilon = 0.01$, $\gamma = 100$.

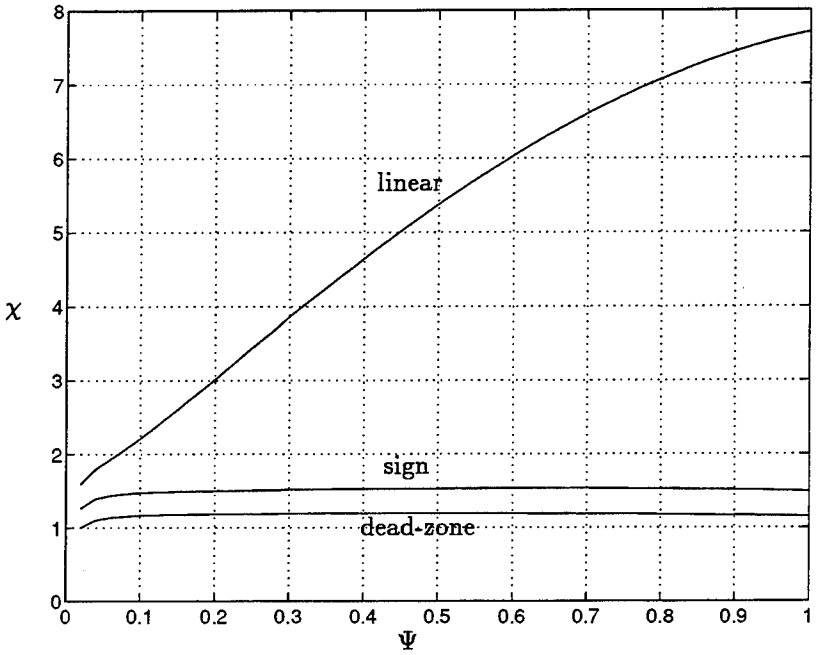

Figure 3.8: Robustness index ($\varepsilon = 0.01$), Gauss-Gauss noise.

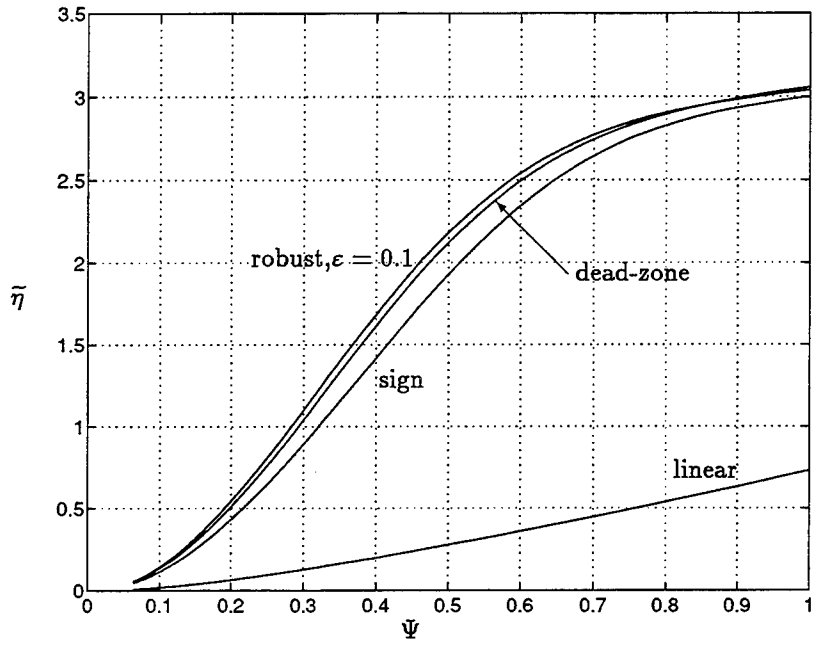

Figure 3.9: $\widetilde{\eta}$ for Gauss-Gauss noise: $\varepsilon = 0.1$, $\gamma = 100$.

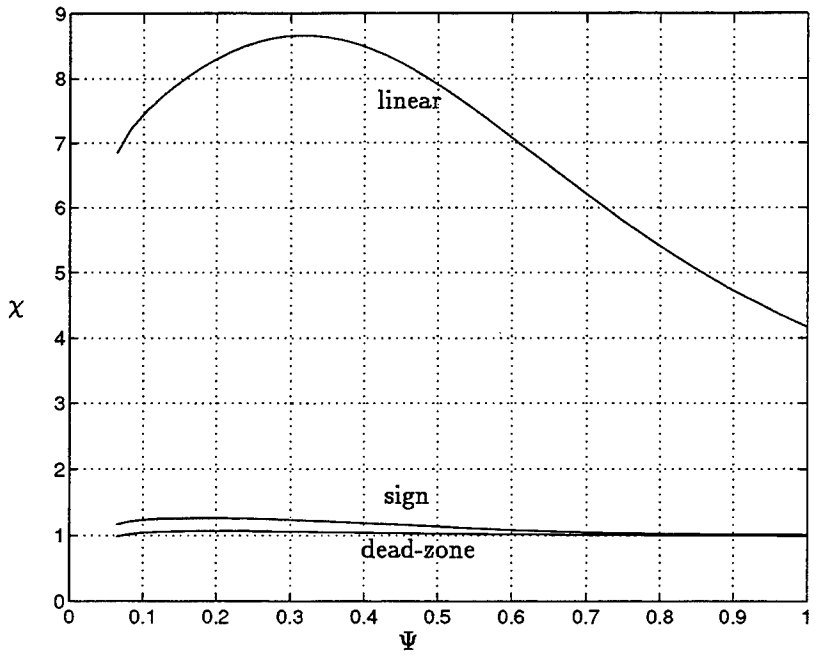

Figure 3.10: Robustness index ($\varepsilon = 0.1$), Gauss-Gauss noise.

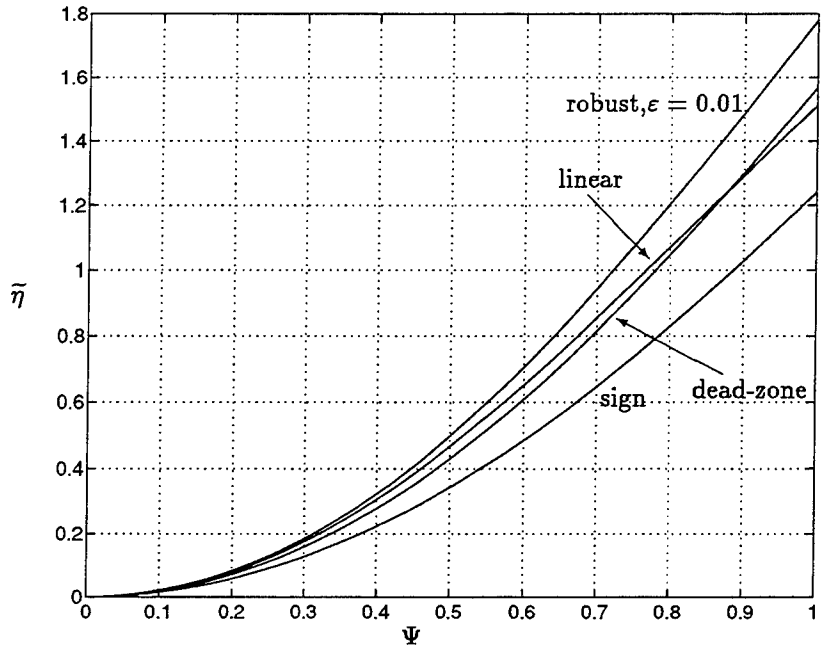

Figure 3.11: $\widetilde{\eta}$ for the least favorable ε-contaminated noise: $\varepsilon = 0.01$.

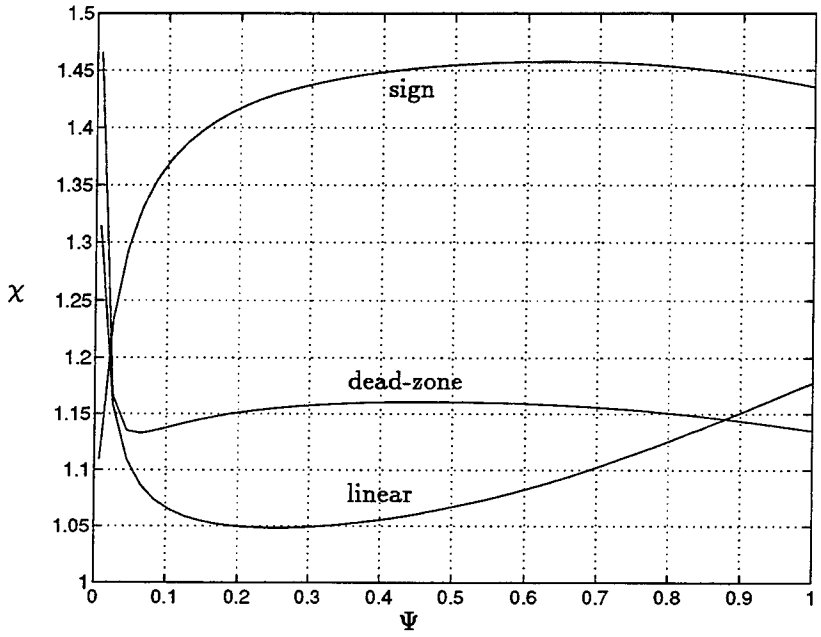

Figure 3.12: Robustness index ($\varepsilon = 0.01$), least favorable ε-contaminated noise.

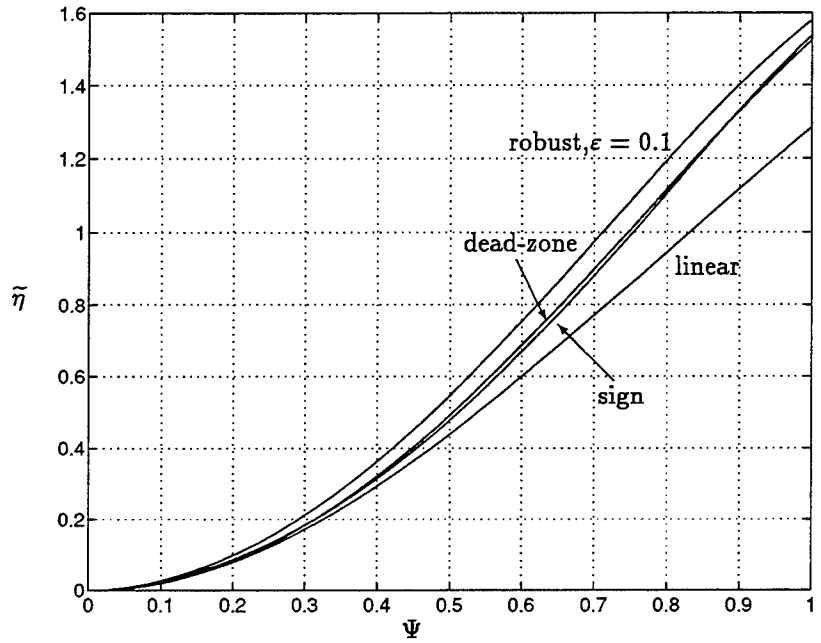

Figure 3.13: $\widetilde{\eta}$ for the least favorable ε-contaminated noise: $\varepsilon = 0.1$.

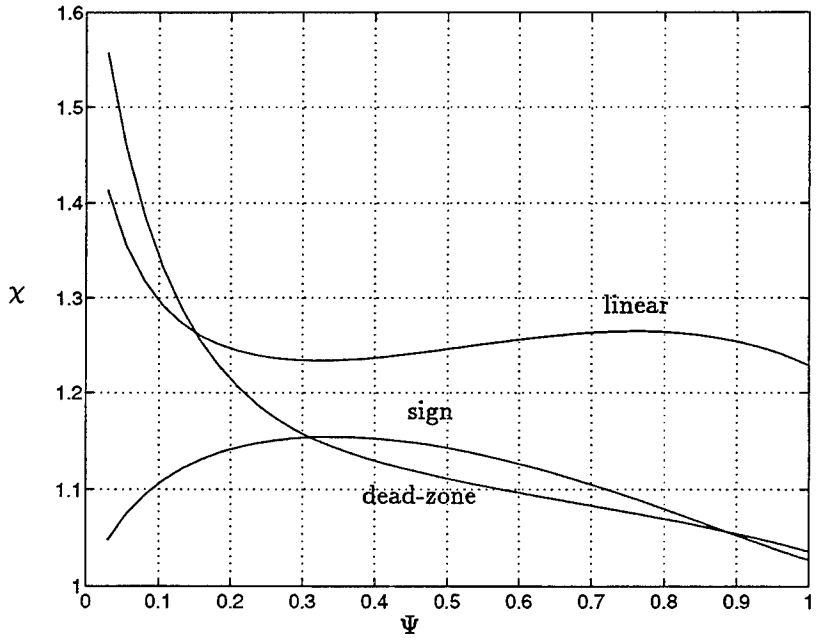

Figure 3.14: Robustness index ($\varepsilon = 0.1$), least favorable ε-contaminated noise.

contamination ($\varepsilon = 0.01$), the linear detector outperforms the nonparametric tests over a fairly large range of Ψ, while requiring less than 10% additional expected delay in most cases (although as Ψ becomes very small, the performance of the linear test deteriorates quickly). On the other hand, observe that one would likely prefer one of the nonparametric tests over the linear test when the contamination is heavier.

Similar conclusions can be drawn when the least favorable total variation noise is present; these results are shown in Figures 3.15 through 3.18. Note that the overall shape of these performance curves bears a striking similarity to those of Figures 3.11 through 3.14. This fact is not surprising when we compare Figure 3.1 and 3.3: the shape of the least favorable densities under the two noise uncertainty models are quite similar, and so one expects that the performance of a given test would be comparable for each.

3.4.2 Illustration of the Saddlepoint Property

Recall that part of the saddlepoint property in (3.4) stated that

$$\eta(g_R; f_{0L}, f_{1L}) = \min_{(f_0,f_1)\in\mathcal{F}_0\times\mathcal{F}_1} \eta(g_R; f_0, f_1)$$

That is, when the robust procedure is used, the minimal performance results when the least favorable distributions are used. This fact is illustrated in Figures 3.19 and 3.20. In each case, we observe that for a fixed uncertainty class (i.e, fixed θ, σ_0, and either ε or δ), the least favorable pair produces the lowest performance. Also notice that as the contamination factors ε and δ increase, the difference between the robust performance for the nominal and least favorable noises increases.

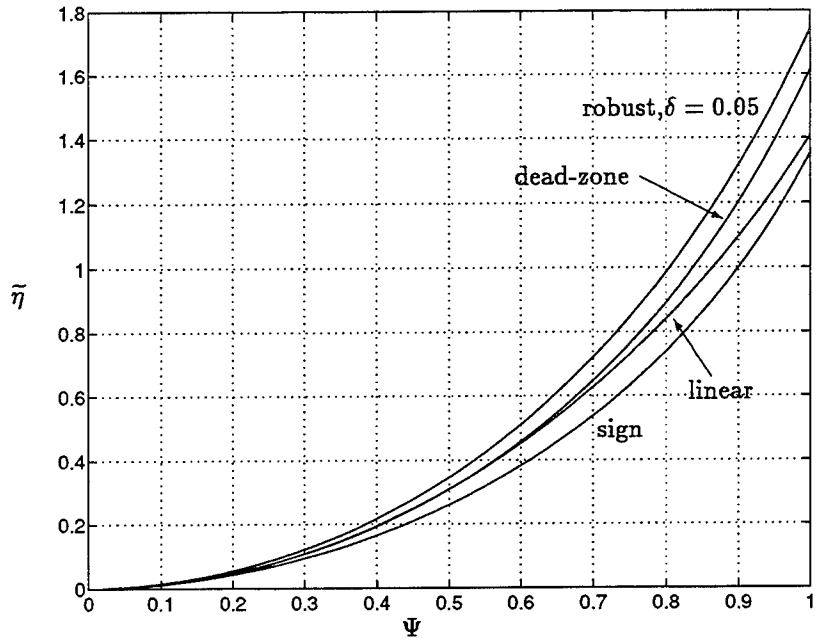

Figure 3.15: $\widetilde{\eta}$ for the least favorable total variation noise: $\delta = 0.05$.

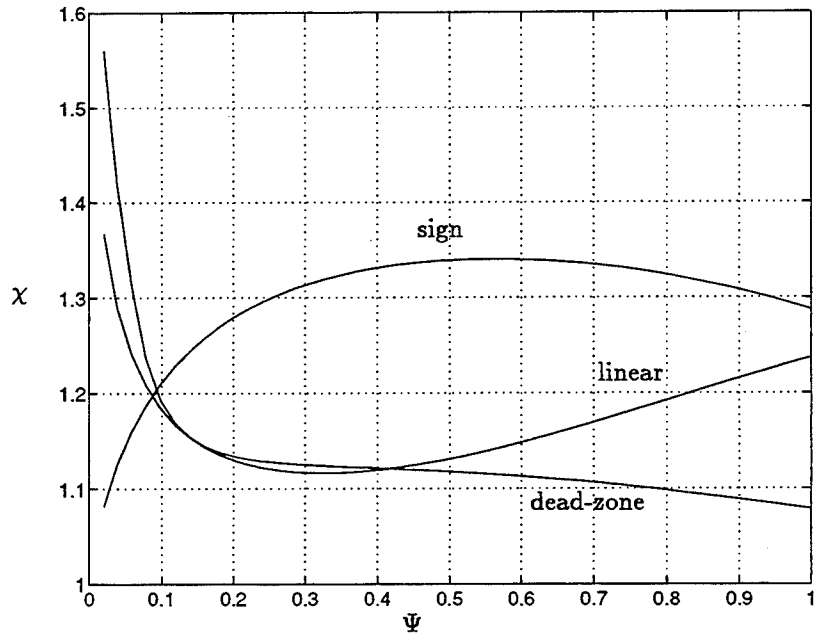

Figure 3.16: Robustness index ($\delta = 0.05$), least favorable total variation noise.

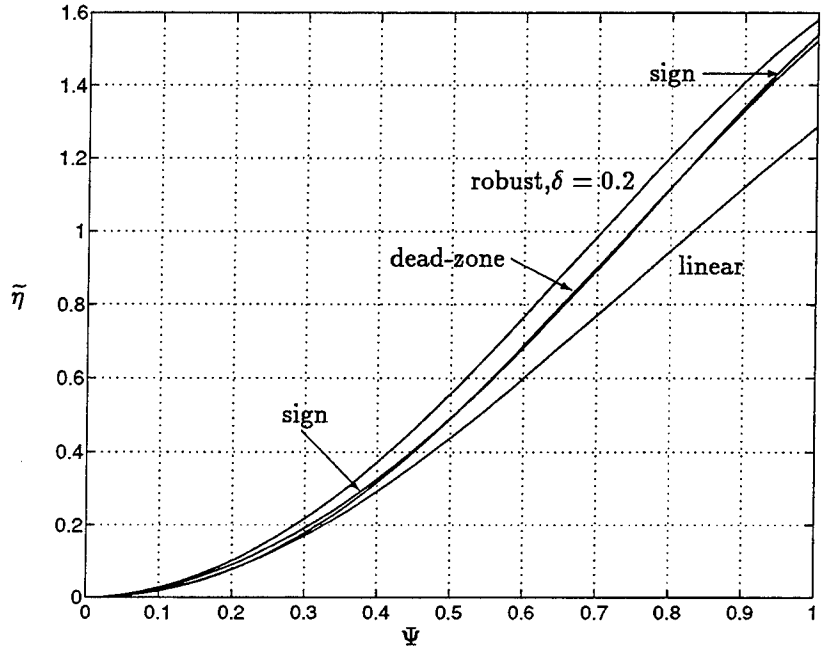

Figure 3.17: $\widetilde{\eta}$ for the least favorable total variation noise: $\delta = 0.2$.

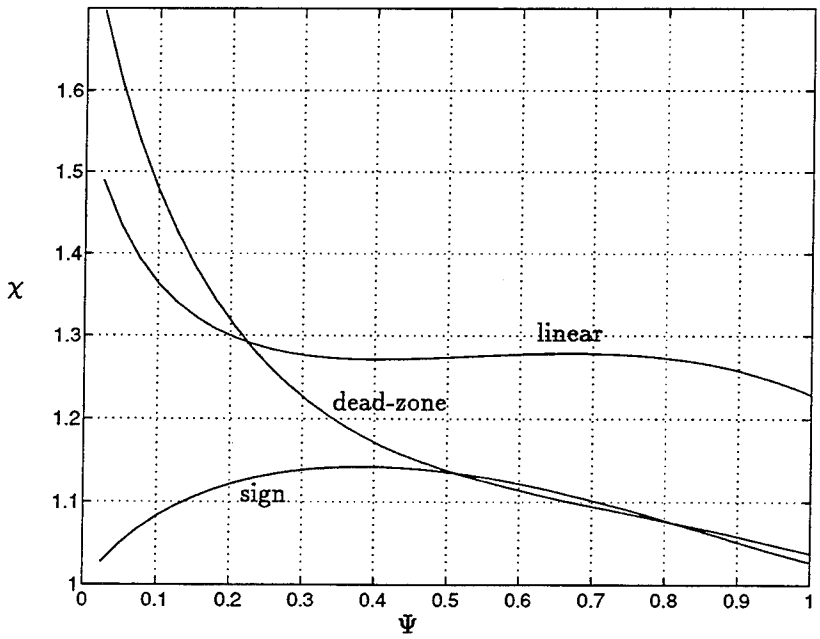

Figure 3.18: Robustness index ($\delta = 0.2$), least favorable total variation noise.

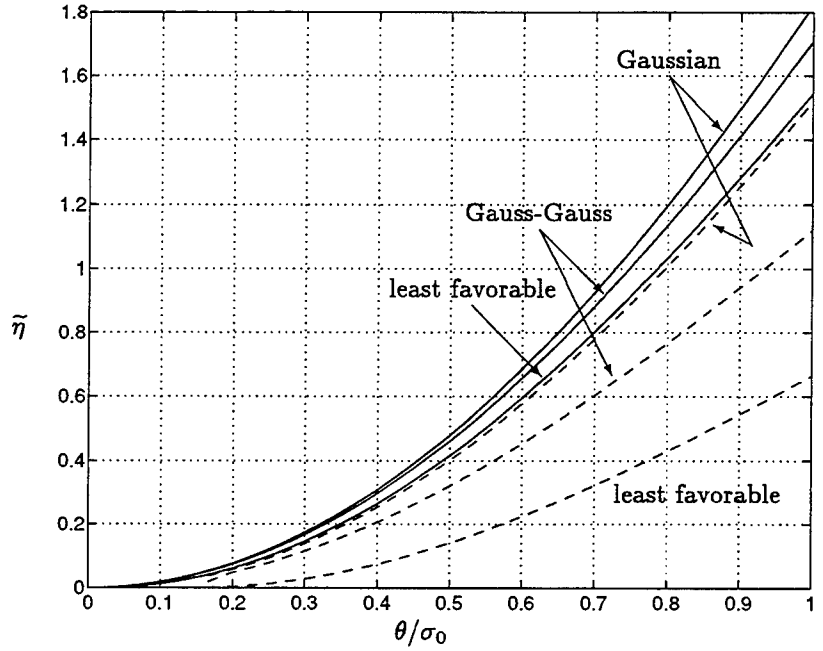

Figure 3.19: $\widetilde{\eta}$ when the robust procedure for the ε-contaminated class is used. The type of noise is indicated on the graph. $\varepsilon = 0.01$ (solid) and 0.1 (dashed).

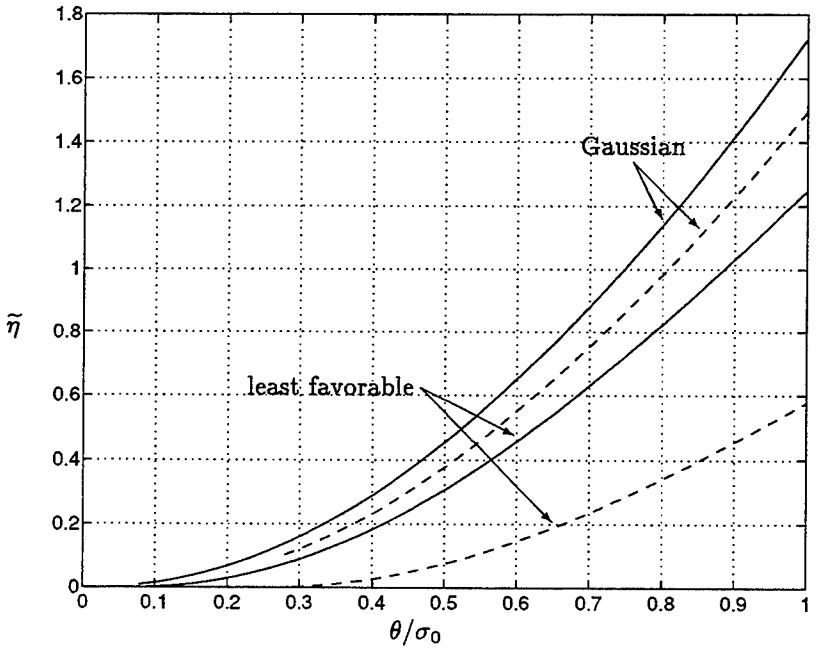

Figure 3.20: $\widetilde{\eta}$ when the robust procedure for the total variation class is used. The type of noise is indicated on the graph. $\delta = 0.05$ (solid) and 0.2 (dashed).

3.4.3 $\tilde{\eta}$ Versus Contamination Level for Different Detector/Noise Combinations

In Figures 3.21 through 3.26, $\tilde{\eta}$ is plotted as a function of the contamination ε and δ for the Gauss-Gauss and least favorable noise types.

It is immediately obvious from Figures 3.21 and 3.22 that the linear detector is a poor choice when Gauss-Gauss noise is present even for very small contamination levels, and regardless of the value of Ψ. Also apparent is that the advantage of the robust detector over the nonparametric procedures is greater when the SNR is lower. However, the dead-zone limiter outperforms the robust test for $\varepsilon > 0.075$ when $\Psi_{dB} = 0$ dB. Finally, in Figure 3.21, notice that for heavy contamination, there is little advantage in opting for the robust procedure over either of the nonparametric tests. Each of these tests shares the property, not possessed by the linear detector, that the observations are processed by "clipping" the larger samples; this property is therefore important when the occurrence of outliers is frequent.

Figures 3.23 and 3.24 show the results for the least favorable noise for the ε-contaminated class. First notice that $\tilde{\eta}$ for the robust and linear detectors are the same for $\varepsilon \to 0$. This is as expected, since for $\varepsilon = 0$, the robust test *is* the linear test (i.e., no robustness is needed, since there is no uncertainty). However, the performance of the linear test falls off quickly, and is the least desirable of all the tests for heavy contamination. Meanwhile, the robust procedure outperforms all of the others for any level of contamination, a fact that results from the saddlepoint property. Again observe that for $\Psi_{dB} = 0$ dB and under heavy contamination ($\varepsilon > 0.1$), there is little advantage of the robust procedure over the nonparametric alternatives.

The results for the least favorable total variation noise are shown in Figures 3.25 and 3.26. As discussed earlier, because the least favorable densities for each of the two class are similar, the resulting performances are also so. Therefore, the same

conclusions can be drawn here as were in the previous discussion.

3.4.4 Example

We conclude this section with an example designed to illustrate the utility of $\tilde{\eta}$. Suppose that one wishes to detect a shift from $-\theta$ to θ in an environment where the noise is not completely characterized; instead, it is known only to be nominally Gaussian, and is otherwise assumed to lie in the ε-contamination class with $\varepsilon = 0.1$ (relatively heavy contamination). It is of interest to design a system which guarantees a maximum rate of false alarms, regardless of what the actual noise distributions turn out to be.

Assume that the observables are of some process which is sampled at 10 kHz, and that $\theta = \sigma_0^2 = 1$. Suppose that it is required to have no more than: *i)* one false alarm per hour, and *ii)* one false alarm per 100 hours (just over 4 days). This results in $T_1 = 3.6 \times 10^7$ samples and $T_2 = 3.6 \times 10^9$ samples. The upper bound on the expected detection delay

$$D_i \leq \frac{\log T_i}{\tilde{\eta}}, \quad i = 1, 2$$

can now be determined simply by computing $\tilde{\eta}$ for the detector and noise distributions of interest (As we observed previously, since $\tilde{\eta} \approx \eta$, the upper bound is actually a good approximation of D).

Table 3 compares the expected delays in milliseconds when using each of the four detectors in Gaussian, Gauss-Gauss, and the least favorable noise with $\varepsilon = 0.1$. Notice that the robust detector outperforms both nonparametric tests in each case, and also that it may not be wise to use the linear detector if there is a large amount of uncertainty in the noise. Furthermore, observe that the additional delay one must incur for raising T by a large amount (from 1 hour to 100 hours) is relatively small since, in the asymptotic realm, the expected delay is proportional to the logarithm

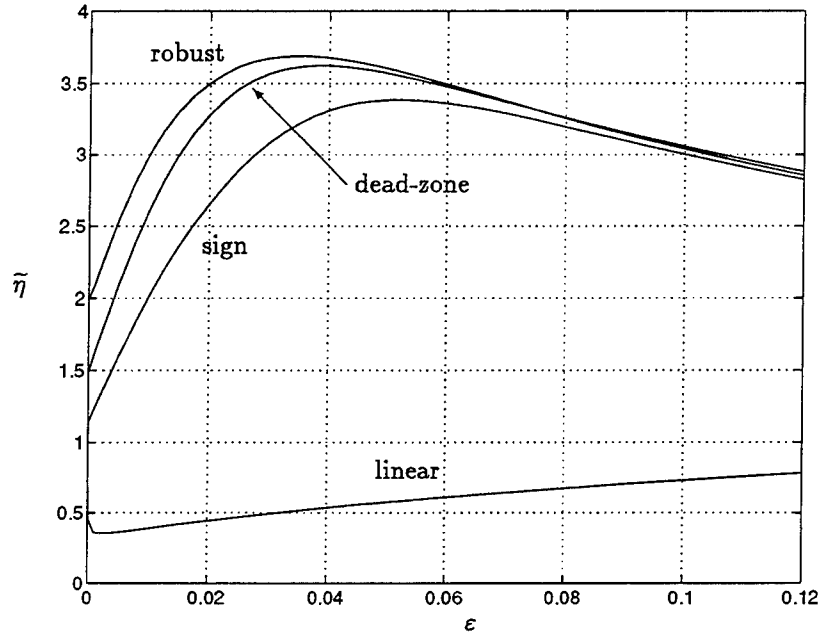

Figure 3.21: $\widetilde{\eta}$ versus ε for Gauss-Gauss noise: $\Psi = 0$ dB, $\gamma = 100$.

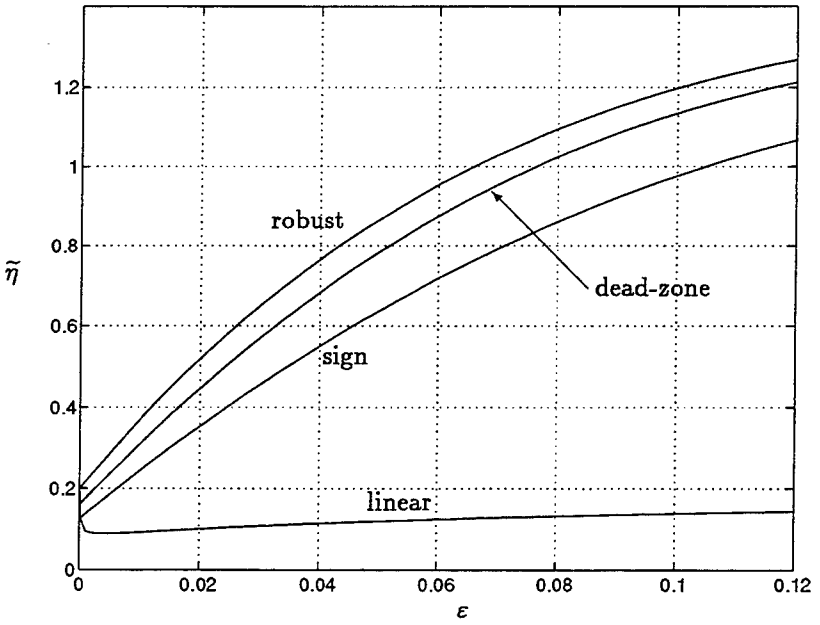

Figure 3.22: $\widetilde{\eta}$ versus ε for Gauss-Gauss noise: $\Psi = -10$ dB, $\gamma = 100$.

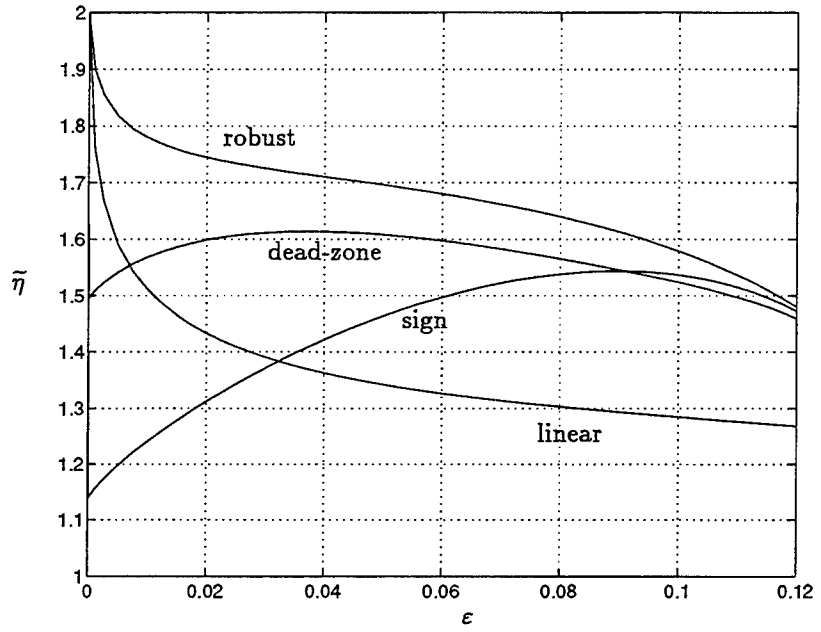

Figure 3.23: $\widetilde{\eta}$ versus ε for least favorable ε-contaminated noise: $\Psi = 0$ dB.

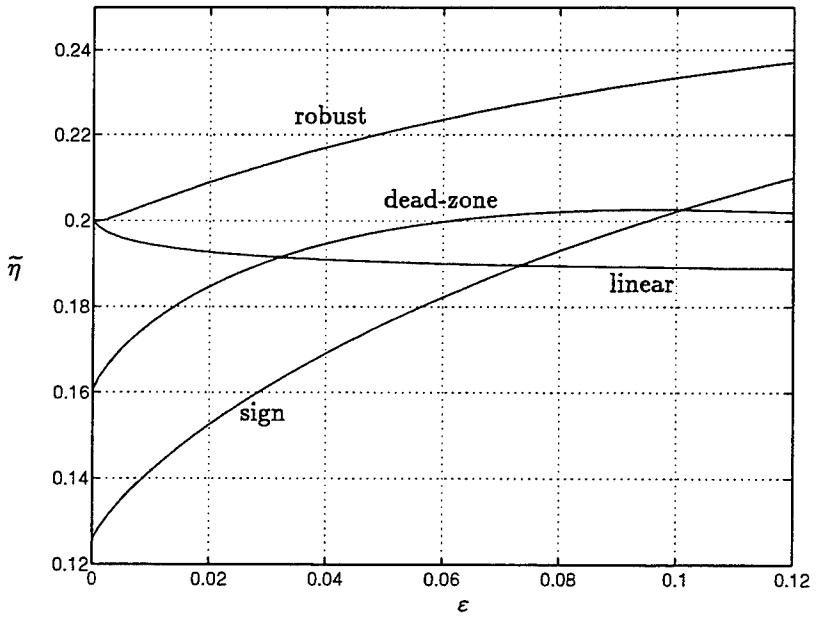

Figure 3.24: $\widetilde{\eta}$ versus ε for least favorable ε-contaminated noise: $\Psi = -10$ dB.

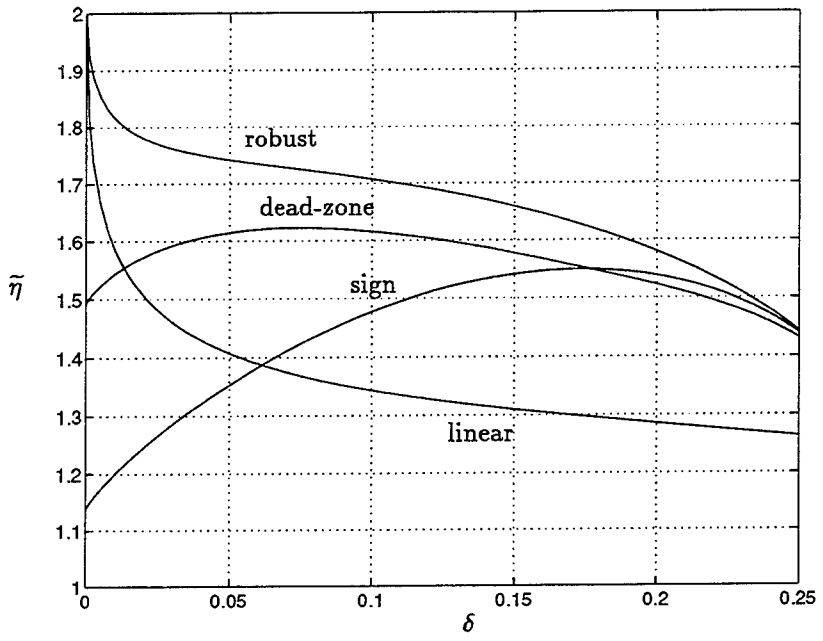

Figure 3.25: $\tilde{\eta}$ versus δ for least favorable total variation noise: $\Psi = 0$ dB.

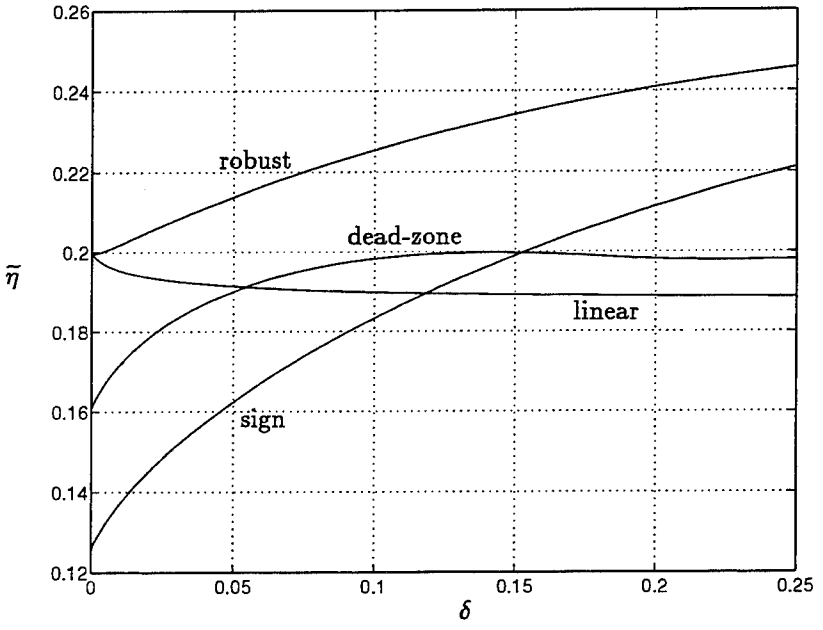

Figure 3.26: $\tilde{\eta}$ versus δ for least favorable total variation noise: $\Psi = -10$ dB.

of T.

Finally, notice the apparent paradox that for the linear test, the expected delay for Gauss-Gauss noise is far greater than that of the "least favorable." Recall that the least favorable densities were those that minimized the asymptotic performance measure when the *optimal* processor, the log-likelihood ratio, was used (see the saddlepoint condition in equation (3.4)). Therefore, the fact that the least favorable noise produces the largest delays is only guaranteed for the robust procedure.

Table 3: Estimate of expected detection delay (ms)

	Gaussian		Gauss-Gauss		Least fav.	
test	T_1	T_2	T_1	T_2	T_1	T_2
linear	0.87	1.10	13.47	17.0	3.28	4.15
sign	1.53	1.93	1.92	2.43	2.97	3.76
dead-zone	1.17	1.47	1.64	2.08	2.89	3.66
robust	1.14	1.45	1.56	1.97	2.62	3.31

3.5 Locally Robust Quickest Detection for the ε-contamination Class

In the preceding sections, the disorder was taken to be a shift from $-\theta$ to θ, where θ was arbitrary. Of particular importance is the case where the disturbance is very small; in other words, θ is close to zero. We are therefore motivated to consider robust quickest detection procedures for the so-called *local* or *weak signal* case.

3.5.1 Local Behavior of the Asymptotic Performance Measure

At the disorder time, suppose the random variables undergo a shift in distribution from $f(x+\theta)$ to $f(x-\theta)$. We have seen that the optimal procedure is Page's test using $g(x) = \log \frac{f(x-\theta)}{f(x+\theta)}$, and that the asymptotic performance is given by the K-L divergence:

$$\tilde{\eta} = \int_{-\infty}^{\infty} \log\left(\frac{f(x-\theta)}{f(x+\theta)}\right) f(x-\theta)dx$$

It is easy to see that $\lim_{\theta\to 0} \tilde{\eta} = 0$. However, this tells us little about the behavior of the performance for weak signals, other than that it becomes increasingly poor as the magnitude of the disturbance approaches zero. Since the *best* asymptotic performance is achieved when g is the log-likelihood ratio, we conclude that the performance will also be poor for arbitrary g.

We want to study the behavior of $\tilde{\eta}$ for θ near zero. A natural approach is to express $\tilde{\eta}$ as a Taylor series about $\theta = 0$ as follows:

$$\tilde{\eta} = \tilde{\eta}|_{\theta=0} + \theta \left.\frac{d\tilde{\eta}}{d\theta}\right|_{\theta=0} + \frac{\theta^2}{2} \left.\frac{d^2\tilde{\eta}}{d\theta^2}\right|_{\theta=0} + \dots$$

Let $\mathcal{G}$ be the collection of all nonlinearities g satisfying $\mathrm{E}\{g(X-\theta) \mid f\} < 0 < \mathrm{E}\{g(X+\theta) \mid f\}$ for all θ in some neighborhood of zero, with $\mathrm{E}\{g(X \pm \theta) \mid f\}$ continuous at $\theta = 0$; then $\mathrm{E}\{g(X) \mid f\} = 0$ for $g \in \mathcal{G}$. It is shown in [2] that

$$\lim_{\theta\to 0} \frac{d\tilde{\eta}}{d\theta} = 0$$

and

$$\lim_{\theta\to 0} \frac{d^2\tilde{\eta}}{d\theta^2} = 4\mathcal{E}$$

where

$$\mathcal{E} = \mathcal{E}(f,g) = \frac{\left(\int g'(x)f(x)dx\right)^2}{\int g^2(x)f(x)dx}$$

is the well-known *efficacy*. The local behavior of $\tilde{\eta}$ when a particular nonlinearity g is used can be approximated by the first nonzero term of the Taylor expansion, which is the second order term, yielding:

$$\tilde{\eta} \approx 2\theta^2 \mathcal{E} \tag{3.20}$$

Therefore, the difference in local asymptotic performance for two different g's can be evaluated by computing the efficacy for each.

3.5.2 Computation of the Locally Robust Quickest Detector

Since the efficacy describes the small signal behavior of Page's test, the minimax criterion can be applied directly to the efficacy in order to obtain the weak signal robust quickest detector. Thus, we need to solve:

$$\max_{g \in \mathcal{G}} \min_{f \in \mathcal{F}} \mathcal{E}(f, g) \tag{3.21}$$

Again, it is useful to determine a saddlepoint solution.

Proposition 3: *There exists a saddlepoint, (g_R, f_L), of (3.21).*

Proof:

It is well-known that for any fixed density $f(x)$, the efficacy is maximized by using

$$g(x) = a\left(\frac{-f'(x)}{f(x)}\right) = a g_{lo}(x), \quad \forall a \neq 0$$

where $g_{lo}(x)$ denotes the locally optimal nonlinearity. A proof using variational calculus is given in [8]. Using this $g(x)$, it is not difficult to show (with the additional condition that $f(-\infty) = f(\infty) = 0$) that

$$\max_{g \in \mathcal{G}} \mathcal{E}(f, g) = \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x) dx = I_F(f)$$

which is just Fisher's information. [10]

Thus, the saddlepoint solution is (g_R, f_L), where $g_R(x) = \frac{-f_L'(x)}{f_L(x)}$ and f_L is the density that minimizes Fisher's information. The existence and uniqueness of such a density are demonstrated in [4]. ∎

Recall that the ε-contaminated noise class with nominal density f_n is

$$\mathcal{F} = \{f(x) : f(x) = (1-\varepsilon)f_n(x) + \varepsilon h(x), \quad h \in \mathcal{H}\}$$

The least favorable distribution (minimizing Fisher's information) is [4]:

$$f_L(x) = \begin{cases} (1-\varepsilon)f_n(x_0)e^{k(x-x_0)}, & x \le x_0 \\ (1-\varepsilon)f_n(x), & x_0 < x < x_1 \\ (1-\varepsilon)f_n(x_1)e^{-k(x-x_1)}, & x \ge x_1 \end{cases}$$

where ε and k are related via

$$\int_{x_0}^{x_1} f_n(x)dx + \frac{f_n(x_0) + f_n(x_1)}{k} = \frac{1}{1-\varepsilon}$$

and where x_0 and x_1 are the endpoints of the interval on which $\left|\frac{f_n'(x)}{f_n(x)}\right| \le k$. The resulting robust nonlinearity is

$$g_R(x) = \frac{-f_L'(x)}{f_L(x)} = \begin{cases} -k, & x \le x_0 \\ \frac{-f_n'(x)}{f_n(x)}, & x_0 < x < x_1 \\ k, & x \ge x_1 \end{cases}$$

3.5.3 The Locally Robust Quickest Detector with Gaussian Nominals

Once again, of particular interest is the case where $f_n(x)$ is the Gaussian density function with zero mean and variance σ_0^2, i.e. $f_n(x) = \varphi(x; \sigma_0)$. Here, notice that

$$\left|\frac{f_n'(x)}{f_n(x)}\right| = \left|\frac{x}{\sigma_0^2}\right| \le k$$

[10] Note that $I_F(f)$ is independent of a. Consequently it is the shape, rather than the scale, of g that determines the local performance.

implies $x_1 = -x_0 = k\sigma_0^2$. The least favorable distribution is now

$$f_L(x) = \begin{cases} \frac{1-\epsilon}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{x^2}{2\sigma_0^2}\right\}, & |\, x \,| \leq k\sigma_0^2 \\ \frac{1-\epsilon}{\sqrt{2\pi}\sigma_0} \exp\left\{\frac{k^2\sigma_0^2}{2} - k\,|x|\right\}, & |\, x \,| \geq k\sigma_0^2 \end{cases}$$

where

$$\frac{2\varphi(k\sigma_0^2; \sigma_0^2)}{k} - 2\Phi(-k\sigma_0) = \frac{\varepsilon}{1-\varepsilon}$$

and

$$g_R(x) = \begin{cases} -k, & x \leq -k\sigma_0^2 \\ \frac{x}{\sigma_0^2}, & |x| < k\sigma_0^2 \\ k, & x \geq k\sigma_0^2 \end{cases}$$

It is interesting to note that the form of this test, a clipped linearity, is the same as that of the large signal test with one exception: the robust weak signal nonlinearity is not a function of the signal strength θ, since it arises under the assumption that $\theta \to 0$.

By using the robust nonlinearity and the least favorable distribution, we can now compute the minimax efficacy, which is just $I_F(f_L)$. We have

$$\begin{aligned} \mathcal{E}(f_L, g_R) &= \int_{\infty}^{\infty} \left(\frac{f_L'(x)}{f_L(x)}\right)^2 f_L(x)dx \\ &= \frac{1-\varepsilon}{\sigma^4} \int_{-k\sigma^2}^{k\sigma^2} \frac{x^2}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} dx + \frac{2k^2(1-\varepsilon)}{\sqrt{2\pi}\sigma} e^{k^2\sigma^2/2} \int_{k\sigma^2}^{\infty} e^{-kx} dx \\ &= \frac{1-\varepsilon}{\sigma^2}\left[1 - 2\Phi(-k\sigma) - 2k\sigma^2\varphi(k\sigma^2; \sigma^2)\right] + \frac{2k(1-\varepsilon)}{\sqrt{2\pi}\sigma} e^{-k^2\sigma^2/2} \\ &= \frac{1}{\sigma^2}\left(1 - \frac{2(1-\varepsilon)(1+k^2\sigma^2)\varphi(k\sigma^2; \sigma^2)}{k}\right) + \frac{2k(1-\varepsilon)}{\sqrt{2\pi}\sigma} e^{-k^2\sigma^2/2} \end{aligned}$$

3.5.4 Comparison of Several Procedures for Local Quickest Detection

The asymptotic performance for other detector and noise combinations can also be obtained by computing the efficacy and then applying (3.20). We can also define

the robustness index χ in the same manner as in Section 4. Suppose we have two procedures, labelled A and B, and let θ be some small positive number close to zero. The performance gain of procedure A relative to procedure B is

$$\chi_{A,B} = \frac{\tilde{\eta}_A}{\tilde{\eta}_B} \approx \frac{2\theta^2 \mathcal{E}_A}{2\theta^2 \mathcal{E}_B} = \frac{\mathcal{E}_A}{\mathcal{E}_B} \tag{3.22}$$

We can recognize the final expression to be the *relative asymptotic efficiency* from the classical hypothesis testing literature. Again, χ can also be interpreted as a measure of the loss or gain in expected delay for one procedure relative to another as described in Section 4.

In Figures 3.27-3.32, the performance for the weak signal scenario is plotted versus ε. Three distributions from the ε-contaminated class are considered: the Gaussian distribution, the Gauss-Gauss mixture, and the least favorable. In each case, the nominal density is Gaussian with zero mean and unit variance.

For each noise type, two plots are given. First, the efficacies resulting from the linear, sign, dead-zone, and robust detectors are computed; the equations for these are given in Appendix E. Second, the performance gain of the robust procedure with respect to the other three is illustrated; that is, $\chi_{A,B}$ is computed as in equation (3.22), where procedure A is the robust procedure, and procedure B is, in turn, each of the other procedures.

In Figures 3.27 and 3.28, we see that for Gaussian noise the robust detector designed assuming a contamination ε outperforms both of the nonparametric tests, but not surprisingly, it is not as effective as the linear test. Nevertheless, the loss in performance in using the robust rather than the linear procedure is less than 10% for $\varepsilon \leq 0.1$, the range of interest in many real world problems.

The results for Gauss-Gauss mixture noise with $\sigma_1^2 = 100$ are shown in Figures 3.29 and 3.30. Again, the robust detector outperforms the nonparametric alternatives. However, notice this time that the performance of the linear detector is very poor

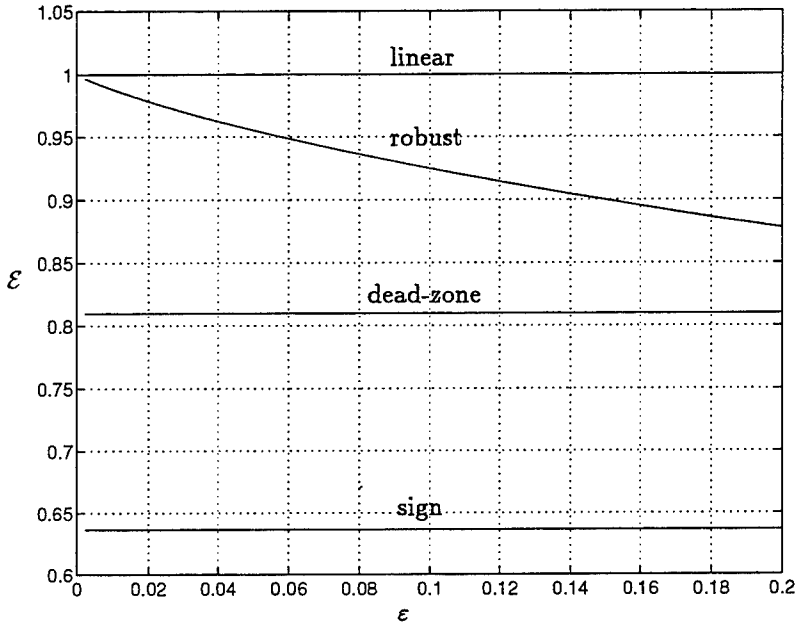

Figure 3.27: Efficacy vs. ε for Gaussian noise.

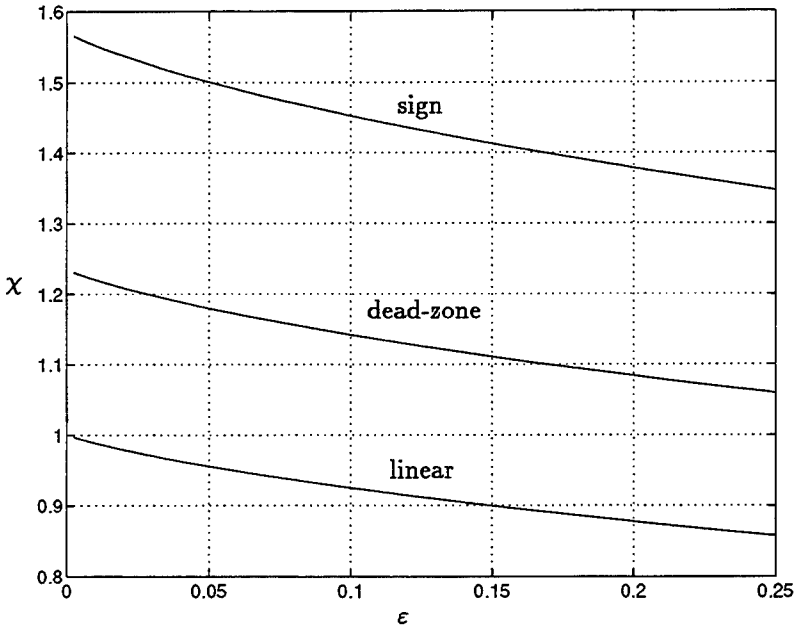

Figure 3.28: Robustness index vs. ε for Gaussian noise.

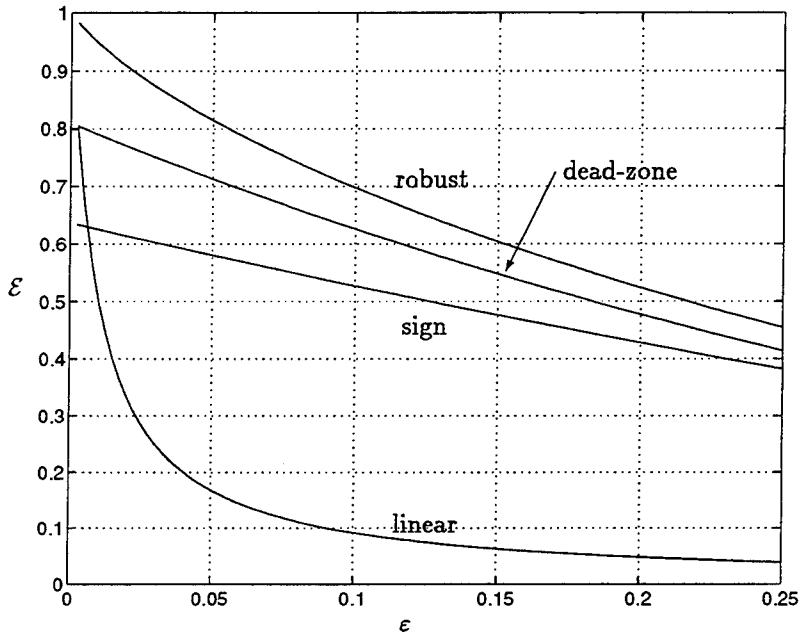

Figure 3.29: Efficacy vs. ϵ for Gauss-Gauss noise, $\gamma = 100$.

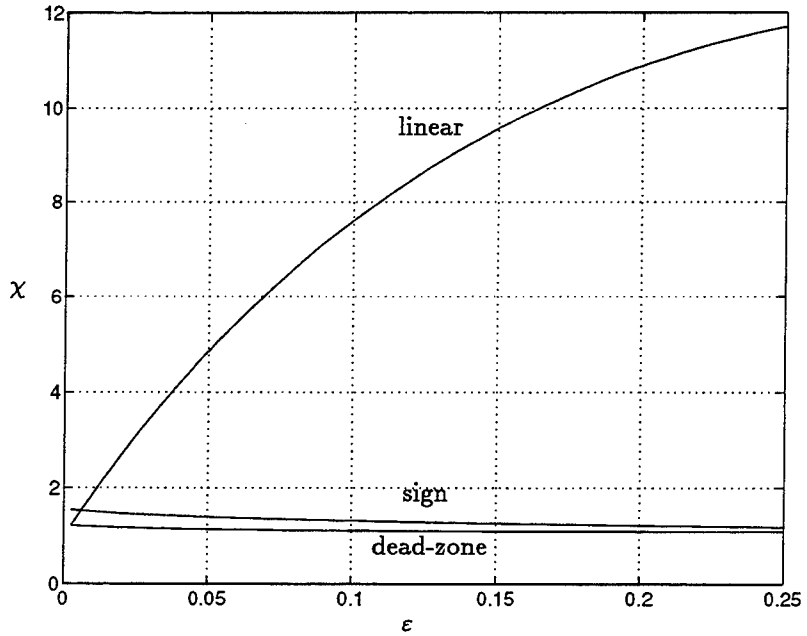

Figure 3.30: Robustness index vs. ϵ for Gauss-Gauss noise, $\gamma = 100$.

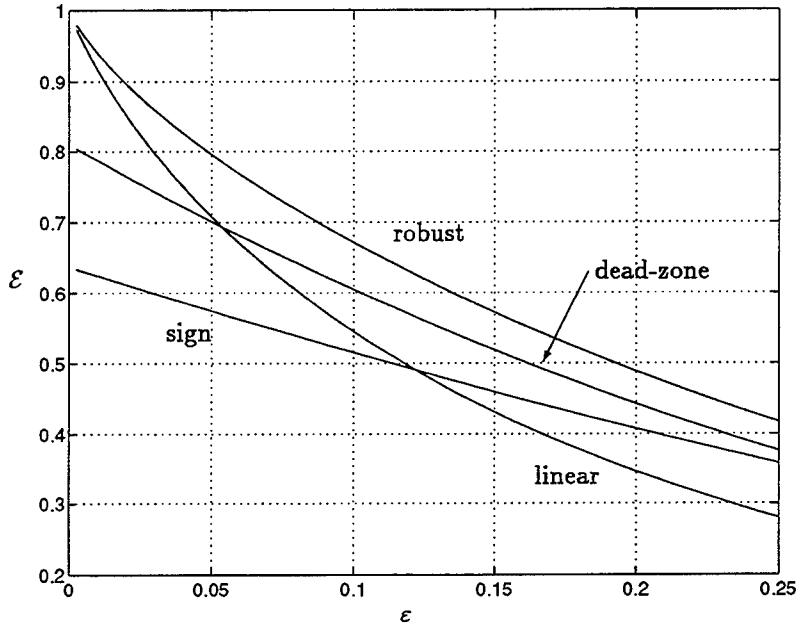

Figure 3.31: Efficacy vs. ϵ for least favorable ϵ-contaminated noise.

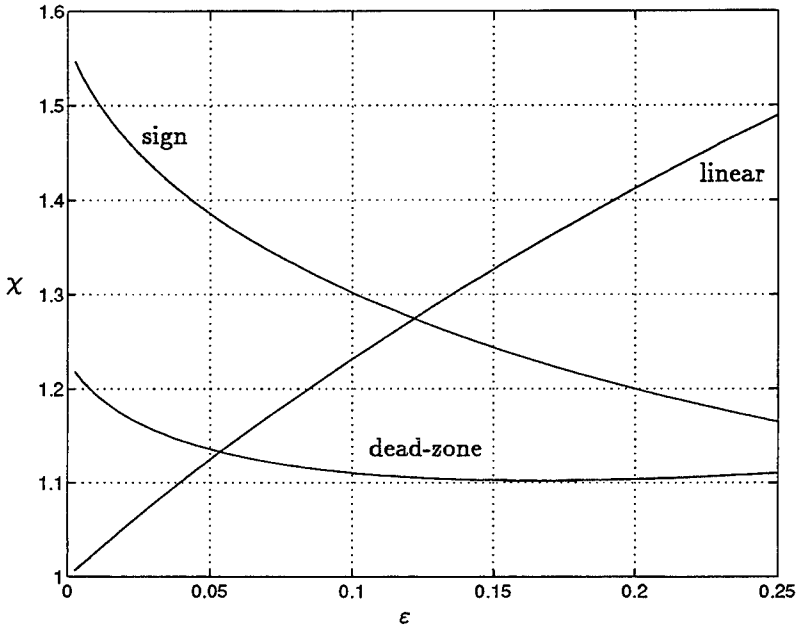

Figure 3.32: Robustness index vs. ϵ for least favorable ϵ-contaminated noise.

when any more than a small amount of contamination is present.

Finally, the performance for the least favorable noise is shown in Figures 3.31 and 3.32. From the the saddlepoint condition of Proposition 2:

$$\max_{g \in \mathcal{G}} \eta(g; f_{0L}, f_{1L}) = \eta(g_R; f_{0L}, f_{1L})$$

and so we expect that the robust detector will outperform each of the others for any ε; the plots confirm this expectation. Notice that for small contamination, the performance of the linear and robust detectors is close. This is not surprising, since for small ε, the least favorable distributions will still be close to the nominals. However, the performance of the linear detector falls off faster than any of the others as ε gets large, and the potential gain in using the robust procedure over the linear detector increases.

3.6 Conclusions

This chapter examines procedures for robust quickest detection when the noise densities are known only to lie within some uncertainty classes, each of which was defined in terms of an allowable deviation from some nominal density. The robust detector was derived by applying the minimax criterion directly to the asymptotic performance measure for Page's test, $\bar{\eta}$, and it was shown that $\bar{\eta}$ exhibits a saddlepoint solution. It was also shown that when the robust processor is used, $\bar{\eta}$ is equal to the Kullback-Leibler divergence, and that the least favorable densities are those that minimize this quantity. Moreover, a formal connection between robust quickest detection and robust hypothesis testing was established, namely the processor used for the former is just the log-likelihood ratio of the least favorable densities in terms of risk. Thus, we were able to apply previous results on robust hypothesis testing to the present problem. This enabled us to obtain the robust quickest detectors for the ε-contaminated and total variation noise uncertainty classes.

The performance of the robust procedure was compared to that of several nonparametric versions of Page's test via the computation of $\bar{\eta}$ versus both SNR and level of uncertainty. It was shown that the robust procedures exhibit good performance over a range of noise distributions within the uncertainty class and outperform the nonparametric alternatives in most cases, yet they are more conservative than the optimal (when the densities are known) procedure. However, procedures which are optimized for the nominal distributions, which in this case were Gaussian, suffered severe degradation in performance in some instances when the true distributions lay elsewhere in the uncertainty class. For example, the procedure using the optimal processor for Gaussian noise, the linear processor, was shown in Section 3.4.1 to produce expected delays which were greater than those of the robust procedure, in some cases by nearly a factor of eight. It was also shown that in situations where a fixed false

alarm rate is desired, the delay to detect the disorder can be maintained at reasonable levels over the entire class if the robust procedure is used.

The locally robust quickest detection procedure was also derived by applying the minimax criterion to the classical efficacy, which is directly proportional to the asymptotic performance in the weak signal scenario; in this case, the least favorable density is that which minimizes Fisher's information. The performance of the locally robust procedure was compared to that of nonparametric alternatives involving the sign detector and dead-zone limiter. The local version of the robust quickest detector exhibited good performance over the entire noise class, outperforming both the linear and nonparametric procedures in most cases.

As a potential area for future work, it would be interesting to examine the performance of the robust procedure when there is a mismatch in the assumed noise class. For example, one might assume that $\varepsilon = \varepsilon_1$, when in fact the contamination is $\varepsilon_2 \neq \varepsilon_1$. One would suspect that for small mismatch, there would be only slight degradation. This is an important area, since the contamination factor is often not known exactly, but estimated from the data. This raises the additional question of how good one's estimate of ε must be in order to design a robust test. Also, the method for deriving the minimax robust procedure via the asymptotic performance measure could be applied to multivariate models. This method does not assume that the process is scalar, and so some results should carry over directly. In Chapter 4, we examine the related, but more specific problem of quickest detection for Gaussian noise with unknown mean vector and covariance matrix.

3.7 Appendices

3.7.A Computation of $\tilde{\eta}$ for Various Noise/Detector Combinations Involving the ε-contaminated Class

The general procedure for computing $\tilde{\eta}$ is given in the text. Below, the performance is computed for some specific cases involving the ε-contaminated noise uncertainty class. Throughout, the procedure is designed to detect a shift in the mean from $-\theta$ to θ, and the nominal distribution of the ε-contaminated class is Gaussian with variance σ_0^2. The constants c_0 and c_1 are chosen to satisfy

$$\begin{aligned} \Phi\left(\frac{b_0+\theta}{\sigma_0}\right) + c_0^{-1}\Phi\left(\frac{-b_0+\theta}{\sigma_0}\right) &= \frac{1}{1-\varepsilon} \\ \Phi\left(\frac{-b_1+\theta}{\sigma_0}\right) + c_1\Phi\left(\frac{b_1+\theta}{\sigma_0}\right) &= \frac{1}{1-\varepsilon} \end{aligned}$$

where $b_0 = \frac{\sigma_0^2}{2\theta}\log c_0$ and $b_1 = \frac{\sigma_0^2}{2\theta}\log c_1$. Note that $c_1 < 1 < c_0$, and so $b_1 < 0 < b_0$.

Gauss-Gauss Noise, Robust Detector for the ε-contaminated Class

In this section, $\tilde{\eta}$ is computed where the robust detector for the ε-contaminated class is used and the noise is Gauss-Gauss with contamination factor ε:

$$f(x) = (1-\varepsilon)\varphi(x;\sigma_0) + \varepsilon\varphi(x;\sigma_1)$$

Here $\varphi(x;\sigma)$ is the Gaussian density with variance σ^2; thus, the noise is nominally Gaussian with variance σ_0^2, but is occasionally (with probability ε) contaminated by impulsive noise modelled as Gaussian with variance $\sigma_1^2 \gg \sigma_0^2$.

Define the following:

$$\begin{array}{ll} f_0(x)\triangleq f(x+\theta) & f_1(x)\triangleq f(x-\theta) \\ \phi_0(x)\triangleq\varphi(x+\theta;\sigma_0) & \phi_1(x)\triangleq\varphi(x-\theta;\sigma_0) \\ \varsigma_0(x)\triangleq\varphi(x+\theta;\sigma_1) & \varsigma_0(x)\triangleq\varphi(x-\theta;\sigma_1) \end{array}$$

The robust nonlinearity is

$$g_R(x) = \log \frac{f_{L1}(x)}{f_{L0}(x)} = \begin{cases} \log c_1, & \frac{\phi_1}{\phi_0} \le c_1 \\ \log \frac{\phi_1(x)}{\phi_0(x)}, & c_1 < \frac{\phi_1}{\phi_0} \le c_0 \\ \log c_0, & \frac{\phi_1}{\phi_0} \ge c_0 \end{cases}$$

Therefore, we are required to compute

$$\tilde{\eta} = \omega_0 \mathrm{E}\{g_R(x) \mid f_1\} \tag{A.1}$$

where ω_0 satisfies

$$\mathrm{E}\{e^{\omega_0 g_R(x)} \mid f_0\} = 1$$

Observe that the linearity of the expectation operator allows us to write:

$$\mathrm{E}\{e^{\omega_0 g_R(x)} \mid f_0\} = (1-\varepsilon)\mathrm{E}\{e^{\omega_0 g_R(x)} \mid \phi_0\} + \varepsilon \mathrm{E}\{e^{\omega_0 g_R(x)} \mid \varsigma_0\} \tag{A.2}$$

and

$$\mathrm{E}\{g_R(x) \mid f_1\} = (1-\varepsilon)\mathrm{E}\{g_R(x) \mid \phi_1\} + \varepsilon \mathrm{E}\{g_R(x) \mid \varsigma_0\} \tag{A.3}$$

To compute $\mathrm{E}\{e^{\omega_0 g_R(x)} \mid f_0\}$, we first compute $\mathrm{E}\{e^{\omega_0 g_R(x)} \mid \varsigma_0\}$. We have

$$\mathrm{E}\{e^{\omega_0 g_R(x)} \mid \varsigma_0\} = \int_{A_1} c_1^{\omega} \varsigma_0(x)dx + \int_{A_2} \left[\frac{\phi_1(x)}{\phi_0(x)}\right]^{\omega} \varsigma_0(x)dx + \int_{A_3} c_0^{\omega} \varsigma_0(x)dx$$

where $A_1 = \{\frac{\phi_1}{\phi_0} \le c_1\} = \{x \le b_1\}$, $A_2 = \{c_1 < \frac{\phi_1}{\phi_0} < c_0\} = \{b_1 \le x \le b_0\}$, $A_3 = \{\frac{\phi_1}{\phi_0} \ge c_0\} = \{x \ge b_0\}$. The integrals over A_1 and A_3 are easy to compute. To compute the A_2 term, first let $\nu = \frac{2\omega\theta}{\sigma_0^2}$; then:

$$\begin{aligned} \int_{A_2} \left[\frac{\phi_1(x)}{\phi_0(x)}\right]^{\omega} \varsigma_0(x)dx &= \int_{b_1}^{b_0} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{\frac{-1}{2\sigma_1^2}(x^2 + 2(\theta - \sigma_1^2\nu)x + \theta^2)\right\} dx \\ &= \exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega\left(1 - \frac{\sigma_1^2\omega}{\sigma_0^2}\right)\right\} \int_{b_1}^{b_0} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{\frac{-1}{2\sigma_1^2}\left(x + \theta\left(1 - \frac{2\sigma_1^2\omega}{\sigma_0^2}\right)\right)^2\right\} dx \\ &= \exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega\left(1 - \frac{\sigma_1^2\omega}{\sigma_0^2}\right)\right\} \left[\Phi\left(\frac{b_0 + \theta(1 - \frac{2\sigma_1^2\omega}{\sigma_0^2})}{\sigma_1}\right) - \Phi\left(\frac{b_1 + \theta(1 - \frac{2\sigma_1^2\omega}{\sigma_0^2})}{\sigma_1}\right)\right] \end{aligned}$$

Combining this with the other two integrals, we have

$$\mathrm{E}\{e^{\omega_0 g_R(x)} \mid \varsigma_0\} = c_1^{\omega}\Phi\left(\frac{b_1+\theta}{\sigma_1}\right)+c_0^{\omega}\Phi\left(\frac{-b_0-\theta}{\sigma_1}\right)$$
$$+ \exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega\left(1-\frac{\sigma_1^2\omega}{\sigma_0^2}\right)\right\}\left[\Phi\left(\frac{b_0+\theta(1-\frac{2\sigma_1^2\omega}{\sigma_0^2})}{\sigma_1}\right)-\Phi\left(\frac{b_1+\theta(1-\frac{2\sigma_1^2\omega}{\sigma_0^2})}{\sigma_1}\right)\right]$$

Also, observe that we can easily obtain $\mathrm{E}\{e^{\omega_0 g_R(x)} \mid \phi_0\}$ by noting that

$$\mathrm{E}\{e^{\omega_0 g_R(x)} \mid \phi_0\} = \mathrm{E}\{e^{\omega_0 g_R(x)} \mid \varsigma_0\}\Big|_{\sigma_1=\sigma_0}$$

Therefore using (A.2), we have that ω_0 is the nonzero root of the equation

$$\begin{aligned}
1 &= (1-\varepsilon)\exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega(1-\omega)\right\}\left[\Phi\left(\frac{b_0+\theta(1-2\omega)}{\sigma_0}\right)-\Phi\left(\frac{b_1+\theta(1-2\omega)}{\sigma_0}\right)\right] \\
&+ \varepsilon\exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega\left(1-\frac{\sigma_1^2\omega}{\sigma_0^2}\right)\right\}\left[\Phi\left(\frac{b_0+\theta(1-\frac{2\sigma_1^2\omega}{\sigma_0^2})}{\sigma_1}\right)-\Phi\left(\frac{b_1+\theta(1-\frac{2\sigma_1^2\omega}{\sigma_0^2})}{\sigma_1}\right)\right] \\
&+ c_1^{\omega}\left[(1-\varepsilon)\Phi\left(\frac{b_1+\theta}{\sigma_0}\right)+\varepsilon\Phi\left(\frac{b_1+\theta}{\sigma_1}\right)\right] \\
&+ c_0^{\omega}\left[(1-\varepsilon)\Phi\left(\frac{-b_0-\theta}{\sigma_0}\right)+\varepsilon\Phi\left(\frac{-b_0-\theta}{\sigma_1}\right)\right] \qquad \text{(A.4)}
\end{aligned}$$

A bisection routine can be used to obtain ω_0, where the search region is restricted to the region where the right hand side of (A.4) is increasing.

$\mathrm{E}\{g_R(x) \mid f_1\}$ is computed similarly by first computing $\mathrm{E}\{g_R(x) \mid \varsigma_1\}$ and then using the fact that

$$\mathrm{E}\{g_R(x) \mid \phi_1\} = \mathrm{E}\{g_R(x) \mid \varsigma_1\}\Big|_{\sigma_1=\sigma_0}$$

The final result is

$$\mathrm{E}\{g_R(x) \mid f_1\} = \log c_1\left[(1-\varepsilon)\Phi\left(\frac{b_1-\theta}{\sigma_0}\right)+\varepsilon\Phi\left(\frac{b_1-\theta}{\sigma_1}\right)\right]$$

$$
\begin{aligned}
&+ \log c_0 \left[(1-\varepsilon)\Phi\left(\frac{-b_0+\theta}{\sigma_0}\right) + \varepsilon\Phi\left(\frac{-b_0+\theta}{\sigma_1}\right)\right] \\
&+ \frac{2\theta^2}{\sigma_0^2}\left\{(1-\varepsilon)\left[\Phi\left(\frac{b_0-\theta}{\sigma_0}\right) - \Phi\left(\frac{b_1-\theta}{\sigma_0}\right)\right] + \varepsilon\left[\Phi\left(\frac{b_0-\theta}{\sigma_1}\right) - \Phi\left(\frac{b_1-\theta}{\sigma_1}\right)\right]\right\} \\
&+ \sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma_0}\left\{(1-\varepsilon)\left[\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\} - \exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}\right]\right. \\
&\left.+ \varepsilon\frac{\sigma_1}{\sigma_0}\left[\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_1^2}\right\} - \exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_1^2}\right\}\right]\right\}
\end{aligned}
$$

Finally, $\tilde{\eta}$ is obtained by using (A.1).

Nominal Gaussian Noise, Robust Detector for the ε-contaminated Class

The asymptotic performance for this case is the same as in the previous section, where we let $\varepsilon = 0$ in the Gauss-Gauss density function. That is, $\tilde{\eta} = \omega_0 \mathrm{E}\{g_R(x) \mid \phi_1\}$ where ω_0 is the unique nonzero root of

$$
\begin{aligned}
1 \;=\; & c_1^{\omega}\Phi\left(\frac{b_1+\theta}{\sigma_0}\right) + c_0^{\omega}\Phi\left(\frac{-b_0-\theta}{\sigma_0}\right) \\
& + \exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega(1-\omega)\right\}\left[\Phi\left(\frac{b_0+\theta(1-2\omega)}{\sigma_0}\right) - \Phi\left(\frac{b_1+\theta(1-2\omega)}{\sigma_0}\right)\right]
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{E}\{g_R(x) \mid f_1\} \;=\; & \log c_1\Phi\left(\frac{b_1-\theta}{\sigma_0}\right) + \log c_0\Phi\left(\frac{-b_0+\theta}{\sigma_0}\right) \\
& + \frac{2\theta^2}{\sigma_0^2}\left[\Phi\left(\frac{b_0-\theta}{\sigma_0}\right) - \Phi\left(\frac{b_1-\theta}{\sigma_0}\right)\right] \\
& + \sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma_0}\left[\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\} - \exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}\right]
\end{aligned}
$$

Gauss-Gauss Noise, Linear Detector

If the noise is assumed to be the nominal Gaussian, the nonlinearity is $g(x) = \frac{2\theta x}{\sigma_0^2}$. However, it was shown in [2] that $\tilde{\eta}$ is invariant to changes in scale. Therefore, for simplicity, we eliminate the constant multiplier and use $g(x) = x$ instead.

For this case, ω_0 is the unique nonzero root of

$$\mathrm{E}\{e^{\omega g(x)} \mid f_0\} = (1-\varepsilon)\mathrm{E}\{e^{\omega x} \mid \phi_0\} + \varepsilon \mathrm{E}\{e^{\omega x} \mid \varsigma_0\}$$

We can alternatively write this equation as

$$\mathrm{E}\{e^{\omega g(x)} \mid f_0\} = (1-\varepsilon)\mathrm{M}_{\phi_0}(\omega) + \varepsilon \mathrm{M}_{\varsigma_0}(\omega)$$

where M_{ϕ_0} and M_{ς_0} are the moment generating functions for the Gaussian densities ϕ_0 and ς_0, respectively, which can be found in a number of references (for example, [10]). Thus, ω_0 is the nonzero root of the equation

$$1 = (1-\varepsilon)\exp\{-\theta\omega + \frac{1}{2}\omega^2\sigma_0^2\} + \varepsilon \exp\{-\theta\omega + \frac{1}{2}\omega^2\sigma_1^2\} \qquad \text{(A.5)}$$

This is a transcendental equation which can be again solved using a bisection routine restricted to a suitable interval. The $\mathrm{E}\{g(x) \mid f_1\}$ term is simply

$$\begin{aligned} \mathrm{E}\{g(x) \mid f\} &= (1-\varepsilon)\mathrm{E}\{x \mid \phi_1\} + \varepsilon \mathrm{E}\{x \mid \varsigma_1\} \\ &= (1-\varepsilon)\theta + \varepsilon\theta = \theta \end{aligned}$$

and so $\tilde{\eta} = \omega_0\theta$ for this case.

Gauss-Gauss Noise, Linear Detector

If the noise were simply Gaussian with variance σ_0^2 ($\varepsilon = 0$), ω_0 in (A.5) could be solved for explicitly, namely $\omega_0 = \frac{2\theta}{\sigma_0^2}$. Note that $\omega_0 > 0$, as expected. The performance is then $\tilde{\eta} = \frac{2\theta^2}{\sigma_0^2}$. Alternatively, one could simply compute the performance directly using the K-L divergence.

Least Favorable ε-contaminated Noise, Linear Detector

The nonlinearity for this procedure is $g(x) = \frac{2\theta x}{\sigma_0^2}$. We are required to compute

$$\tilde{\eta} = \omega_0 \mathrm{E}\{g(x) \mid f_{L1}\}$$

where ω_0 is the non-zero root of the moment generating function equality

$$\mathrm{E}\left\{e^{\omega_0 g(x)} \mid f_{L0}\right\} = 1 \tag{A.6}$$

Recall that

$$f_{L0}(x) = \begin{cases} (1-\varepsilon)\phi_0(x), & \frac{\phi_1}{\phi_0} < c_0 \\ (1-\varepsilon)c_0^{-1}\phi_1(x), & \frac{\phi_1}{\phi_0} \geq c_0 \end{cases}$$

and

$$f_{L1}(x) = \begin{cases} (1-\varepsilon)\phi_1(x), & \frac{\phi_1}{\phi_0} > c_1 \\ (1-\varepsilon)c_1\phi_0(x), & \frac{\phi_1}{\phi_0} \leq c_1 \end{cases}$$

The left side of (A.6) is

$$(1-\varepsilon)\left[\int_{-\infty}^{b_0} e^{\nu x} f_0(x)dx + \frac{1}{c_0}\int_{b_0}^{\infty} e^{\nu x} f_1(x)dx\right]$$

where $\nu \triangleq \frac{2\omega\theta}{\sigma_0^2}$. We will solve the two integrals separately (in each case, this is done by completing the square). The first one is

$$\begin{aligned}
\int_{-\infty}^{b_0} e^{\nu x} f_0(x)dx &= \int_{-\infty}^{b_0} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}\left[x^2 + 2(\theta - \sigma_0^2\nu)x + \theta^2\right]\right\} dx \\
&= \exp\left\{\frac{-2\theta\sigma_0^2\nu + \sigma_0^4\nu^2}{2\sigma_0^2}\right\} \int_{-\infty}^{b_0} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}\left(x + (\theta - \sigma_0^2\nu)\right)^2\right\} dx \\
&= \exp\left\{-\theta\nu + \frac{1}{2}\sigma_0^2\nu^2\right\} \Phi\left(\frac{b_0 + (\theta - \sigma_0^2\nu)}{\sigma_0}\right) \\
&= \exp\left\{\frac{2\theta^2}{\sigma_0^2}\omega(\omega-1)\right\} \Phi\left(\frac{b_0 + \theta(1-2\omega)}{\sigma_0}\right)
\end{aligned}$$

The second integral is derived in a similar manner. We have:

$$\int_{b_0}^{\infty} e^{\nu x} f_1(x)dx = \exp\left\{\frac{2\theta^2}{\sigma_0^2}\omega(\omega+1)\right\} \Phi\left(\frac{-b_0 + \theta(1+2\omega)}{\sigma_0}\right)$$

Thus, the value of ω_0 is determined by obtaining the nonzero root of

$$\begin{aligned}
(1-\varepsilon)\exp\left\{\frac{2\theta^2}{\sigma_0^2}\omega^2\right\} \Bigg[&\exp\{-\frac{2\theta^2}{\sigma_0^2}\omega\} \Phi\left(\frac{b_0 + \theta(1-2\omega)}{\sigma_0}\right) \\
&+ \frac{1}{c_0}\exp\left\{\frac{2\theta^2}{\sigma_0^2}\omega\right\} \Phi\left(\frac{-b_0 + \theta(1+2\omega)}{\sigma_0}\right)\Bigg] = 1
\end{aligned}$$

The term $\mathrm{E}\{g(x) \mid f_{L1}\}$ is given by

$$(1-\varepsilon)\left[\frac{2\theta}{\sigma_0^2}\int_{b_1}^{\infty} x f_1(x)dx + c_1\frac{2\theta}{\sigma_0^2}\int_{-\infty}^{b_1} x f_0(x)dx\right]$$

Compute the two integrals separately. The first is:

$$\begin{aligned}
&\frac{2\theta}{\sigma_0^2}\int_{b_1}^{\infty}\frac{x}{\sqrt{2\pi}\sigma_0}\exp\left\{\frac{-(x-\theta)^2}{2\sigma_0^2}\right\}dx \\
= \;& \frac{2\theta^2}{\sigma_0^2}\int_{b_1-\theta}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{\frac{-y^2}{2\sigma_0^2}\right\}dy + \frac{2\theta}{\sigma_0^2}\int_{b_1-\theta}^{\infty}\frac{y}{\sqrt{2\pi}\sigma_0}\exp\left\{\frac{-y^2}{2\sigma_0^2}\right\}dy \\
= \;& \frac{2\theta^2}{\sigma_0^2}\Phi\left(\frac{-b_1+\theta}{\sigma_0}\right) + \sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma_0}\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\}
\end{aligned}$$

where the change of variables $y = x - \theta$ was used. Similarly, the second integral is

$$-\frac{2\theta^2}{\sigma_0^2}\Phi\left(\frac{b_1+\theta}{\sigma_0}\right) - \sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma_0}\exp\left\{\frac{-(b_1+\theta)^2}{2\sigma_0^2}\right\}$$

Combining the two, we have

$$\begin{aligned}
\mathrm{E}\{g(x) \mid f_1^*\} = \;& (1-\varepsilon)\left\{\frac{2\theta^2}{\sigma_0^2}\left[\Phi\left(\frac{-b_1+\theta}{\sigma_0}\right) - c_1\Phi\left(\frac{b_1+\theta}{\sigma_0}\right)\right]\right. \\
& \left. + \sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma_0}\left[\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\} - c_1\exp\left\{\frac{-(b_1+\theta)^2}{2\sigma_0^2}\right\}\right]\right\}
\end{aligned}$$

$\tilde{\eta}$ can now be computed.

3.7.B Computation of $\tilde{\eta}$ for Various Noise/Detector Combinations Involving the Total Variation Class

The general procedure for computing $\tilde{\eta}$ is given in the text. Below, the performance is computed for some specific cases involving the total variation noise uncertainty model. Throughout, the procedure is designed to detect a shift in the mean from $-\theta$ to θ, and the nominal distribution of the total variation class is Gaussian with

variance σ_0^2. The constants c_0 and c_1 are chosen to satisfy

$$\begin{aligned} \frac{c_0}{1+c_0}\Phi\left(\frac{b_0+\theta}{\sigma_0}\right) - \frac{1}{1+c_0}\Phi\left(\frac{b_0-\theta}{\sigma_0}\right) &= \frac{\delta}{2} \\ -\frac{c_1}{1+c_1}\Phi\left(\frac{-b_1-\theta}{\sigma_0}\right) + \frac{1}{1+c_1}\Phi\left(\frac{-b_1+\theta}{\sigma_0}\right) &= \frac{\delta}{2} \end{aligned}$$

where $b_0 = \frac{\sigma_0^2}{2\theta}\log c_0$ and $b_1 = \frac{\sigma_0^2}{2\theta}\log c_1$. Note that $c_0 < 1 < c_1$, and so $b_0 < 0 < b_1$.

Gaussian Noise, Robust Detector for the Total Variation Class

In this section, $\tilde{\eta}$ is computed where the robust detector for the total variation class is used and the noise is Gaussian with variance σ_0^2.

Let $\varphi(x;\sigma)$ denote the Gaussian density with variance σ^2, and define the following:

$$\phi_0(x) = \varphi(x+\theta;\sigma_0) \quad \phi_1(x) = \varphi(x-\theta;\sigma_0)$$

The robust nonlinearity is

$$g_R(x) = \log\frac{f_{L1}(x)}{f_{L0}(x)} = \begin{cases} \log c_0, & \frac{\phi_1}{\phi_0} \le c_0 \\ \log\frac{\phi_1(x)}{\phi_0(x)}, & c_0 < \frac{\phi_1}{\phi_0} \le c_1 \\ \log c_1, & \frac{\phi_1}{\phi_0} \ge c_1 \end{cases}$$

We are required to compute

$$\tilde{\eta} = \omega_0 \mathrm{E}\{g_R(x) \mid \phi_1\} \tag{B.1}$$

where ω_0 satisfies

$$\mathrm{E}\{e^{\omega_0 g_R(x)} \mid \phi_0\} = 1$$

First we need to compute $\mathrm{E}\{e^{\omega_0 g_R(x)} \mid \phi_0\}$. However, notice that this has already been obtained in Section B.1.1., with the exception that c_0 and c_1 (and also b_0 and b_1) are switched. Therefore ω_0 is the nonzero root of the equation

$$\begin{aligned} 1 \;=\; & c_0^{\omega}\Phi\left(\frac{b_0+\theta}{\sigma_0}\right) + c_1^{\omega}\Phi\left(\frac{-b_1-\theta}{\sigma_0}\right) \\ & + \exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega(1-\omega)\right\}\left[\Phi\left(\frac{b_1+\theta(1-2\omega)}{\sigma_0}\right) - \Phi\left(\frac{b_0+\theta(1-2\omega)}{\sigma_0}\right)\right] \end{aligned} \tag{B.2}$$

A simple bisection routine can be used to obtain ω_0, where the search region is restricted to the region where the right hand side of (B.2) is increasing.

The expression for $\mathrm{E}\{g_R(x) \mid \phi_1\}$ was also obtained in Section B.1.1. with the aforementioned constants swapped. The result for the present case is

$$\begin{aligned}\mathrm{E}\{g_R(x) \mid \phi_1\} &= \log c_0 \Phi\left(\frac{b_0-\theta}{\sigma_0}\right) + \log c_1 \Phi\left(\frac{-b_1+\theta}{\sigma_0}\right) \\ &\quad + \frac{2\theta^2}{\sigma_0^2}\left[\Phi\left(\frac{b_1-\theta}{\sigma_0}\right) - \Phi\left(\frac{b_0-\theta}{\sigma_0}\right)\right] \\ &\quad + \sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma_0}\left[\exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\} - \exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\}\right]\end{aligned}$$

Finally, $\tilde{\eta}$ is obtained by using (B.1).

Least Favorable Total Variation Noise, Linear Detector

The nonlinearity for this procedure is $g(x) = \frac{2\theta x}{\sigma_0^2}$. We are required to compute

$$\tilde{\eta} = \omega_0 \mathrm{E}\{g(x) \mid f_{L1}\}$$

where ω_0 is the non-zero root of the moment generating function equality

$$\mathrm{E}\left\{e^{\omega_0 g(x)} \mid f_{L0}\right\} = 1 \tag{B.3}$$

Recall that

$$f_{L0}(x) = \begin{cases} \frac{1}{1+c_0}(\phi_0(x)+\phi(x)), & \frac{\phi_1}{\phi_0} \le c_0 \\ \phi_0(x), & c_0 < \frac{\phi_1}{\phi_0} < c_1 \\ \frac{1}{1+c_1}(\phi_0(x)+\phi(x)), & \frac{\phi_1}{\phi_0} \ge c_1 \end{cases}$$

and

$$f_{L1}(x) = \begin{cases} \frac{c_0}{1+c_0}(\phi_0(x)+\phi(x)), & \frac{\phi_1}{\phi_0} \le c_0 \\ \phi_1(x), & c_0 < \frac{\phi_1}{\phi_0} < c_1 \\ \frac{c_1}{1+c_1}(\phi_0(x)+\phi(x)), & \frac{\phi_1}{\phi_0} \ge c_1 \end{cases}$$

The left side of (B.3) is

$$\frac{1}{1+c_0}\int_{-\infty}^{b_0} e^{\nu x}(\phi_0(x)+\phi_1(x))dx+\int_{b_0}^{b_1} e^{\nu x}\phi_0(x)dx+\frac{1}{1+c_1}\int_{b_1}^{\infty} e^{\nu x}(\phi_0(x)+\phi_1(x))dx$$

These integrals can be solved by collecting the powers of the exponentials and completing the square. The result is that ω_0 satisfies

$$\begin{aligned}
1 \;=\; & \frac{1}{1+c_0}\left[\exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega(1-\omega)\right\}\Phi\left(\frac{b_0+\theta(1-2\omega)}{\sigma_0}\right)\right. \\
& \left.+\exp\left\{\frac{2\theta^2}{\sigma_0^2}\omega(1+\omega)\right\}\Phi\left(\frac{b_0-\theta(1+2\omega)}{\sigma_0}\right)\right] \\
& +\frac{1}{1+c_1}\left[\exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega(1-\omega)\right\}\Phi\left(\frac{-b_1-\theta(1-2\omega)}{\sigma_0}\right)\right. \\
& \left.+\exp\left\{\frac{2\theta^2}{\sigma_0^2}\omega(1+\omega)\right\}\Phi\left(\frac{-b_1+\theta(1+2\omega)}{\sigma_0}\right)\right] \\
& +\exp\left\{\frac{-2\theta^2}{\sigma_0^2}\omega(1-\omega)\right\}\left[\Phi\left(\frac{b_1+\theta(1-2\omega)}{\sigma_0}\right)-\Phi\left(\frac{b_0+\theta(1-2\omega)}{\sigma_0}\right)\right]
\end{aligned}$$

The term $\mathrm{E}\{g(x) \mid f_{L1}\}$ is given by

$$\frac{c_0}{1+c_0}\int_{-\infty}^{b_0}\frac{2\theta x}{\sigma_0^2}\left(\phi_0(x)+\phi_1(x)\right)dx+\int_{b_0}^{b_1}\frac{2\theta x}{\sigma_0^2}\phi_1(x)dx+\frac{c_1}{1+c_1}\int_{b_1}^{\infty}\frac{2\theta x}{\sigma_0^2}\left(\phi_0(x)+\phi_1(x)\right)dx$$

These integrals can be solved via a change of variables followed by integration by parts. The result is

$$\begin{aligned}
\mathrm{E}\{g(x) \mid f_{L1}\} \;=\; & \frac{c_0}{1+c_0}\left\{-\sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma}\left[\exp\left\{\frac{-(b_0+\theta)^2}{2\sigma_0^2}\right\}+\exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}\right]\right. \\
& \left.+\frac{2\theta^2}{\sigma_0^2}\left[\Phi\left(\frac{b_0-\theta}{\sigma_0}\right)-\Phi\left(\frac{b_0+\theta}{\sigma_0}\right)\right]\right\} \\
& +\frac{c_1}{1+c_1}\left\{\sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma}\left[\exp\left\{\frac{-(b_1+\theta)^2}{2\sigma_0^2}\right\}+\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\}\right]\right. \\
& \left.+\frac{2\theta^2}{\sigma_0^2}\left[\Phi\left(\frac{-b_1+\theta}{\sigma_0}\right)-\Phi\left(\frac{-b_1-\theta}{\sigma_0}\right)\right]\right\} \\
& +\sqrt{\frac{2}{\pi}}\frac{\theta}{\sigma}\left[\exp\left\{\frac{-(b_0+\theta)^2}{2\sigma_0^2}\right\}-\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\}\right] \\
& +\frac{2\theta^2}{\sigma_0^2}\left[\Phi\left(\frac{b_1-\theta}{\sigma_0}\right)-\Phi\left(\frac{b_0-\theta}{\sigma_0}\right)\right]
\end{aligned}$$

and $\tilde{\eta}$ can be obtained directly.

3.7.C Variance of the Least Favorable Distributions

In Section 4 the effective signal to noise ratio was defined as

$$\Psi = \frac{\mathrm{E}\{X \mid f_1\}}{\sqrt{\mathrm{Var}\{X \mid f_1\}}}$$

The variance of the least favorable distributions is simply given by

$$\mathrm{Var}\{X \mid f_{L1}\} = \mathrm{E}\{X^2 \mid f_{L1}\} - (\mathrm{E}\{X \mid f_{L1}\})^2$$

The expressions for $\mathrm{E}\{X \mid f_{L1}\}$ and $\mathrm{E}\{X^2 \mid f_{L1}\}$ are given below for each of the two noise classes.

ε-contamination Model

$$\begin{aligned}
\mathrm{E}\{X \mid f_{L1}\} &= \int_{-\infty}^{\infty} x f_{L1}(x)dx \\
&= \frac{(1-\varepsilon)\sigma_0}{\sqrt{2\pi}}\left[\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\} - c_1 \exp\left\{\frac{-(b_1+\theta)^2}{2\sigma_0^2}\right\}\right] \\
&+ \theta(1-\varepsilon)\left[\Phi\left(\frac{-b_1+\theta}{\sigma_0}\right) - c_1\Phi\left(\frac{b_1+\theta}{\sigma_0}\right)\right]
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\{X^2 \mid f_{L1}\} &= \int_{-\infty}^{\infty} x^2 f_{L1}(x)dx \\
&= c_1(1-\varepsilon)\left\{\frac{\sigma_0(\theta-b_1)}{\sqrt{2\pi}}\exp\left\{\frac{-(b_1+\theta)^2}{2\sigma_0^2}\right\} + (\theta^2+\sigma_0^2)\Phi\left(\frac{b_1+\theta}{\sigma_0}\right)\right\} \\
&+ (1-\varepsilon)\left\{\frac{\sigma_0(\theta+b_1)}{\sqrt{2\pi}}\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\} + (\theta^2+\sigma_0^2)\Phi\left(\frac{-b_0+\theta}{\sigma_0}\right)\right\}
\end{aligned}$$

Total Variation Model

$$
\begin{aligned}
\mathrm{E}\{X \mid f_{L1}\} &= \int_{-\infty}^{\infty} x f_{L1}(x)dx \\
&= \frac{c_0}{1+c_0}\left\{\theta\left[\Phi\left(\frac{b_0-\theta}{\sigma_0}\right)-\Phi\left(\frac{b_0+\theta}{\sigma_0}\right)\right]\right. \\
&- \left.\frac{\sigma_0}{2\pi}\left[\exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}+\exp\left\{\frac{-(b_0+\theta)^2}{2\sigma_0^2}\right\}\right]\right\} \\
&+ \frac{c_1}{1+c_1}\left\{\theta\left[\Phi\left(\frac{-b_1+\theta}{\sigma_0}\right)-\Phi\left(\frac{-b_1-\theta}{\sigma_0}\right)\right]\right. \\
&+ \left.\frac{\sigma_0}{2\pi}\left[\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\}+\exp\left\{\frac{-(b_1+\theta)^2}{2\sigma_0^2}\right\}\right]\right\} \\
&+ \theta\left[\Phi\left(\frac{b_1-\theta}{\sigma_0}\right)-\Phi\left(\frac{b_0-\theta}{\sigma_0}\right)\right] \\
&+ \frac{\sigma_0}{2\pi}\left[\exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}-\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\}\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\{X^2 \mid f_{L1}\} &= \int_{-\infty}^{\infty} x^2 f_{L1}(x)dx \\
&= \frac{c_0}{1+c_0}\left\{(\theta^2+\sigma_0^2)\left[\Phi\left(\frac{b_0+\theta}{\sigma_0}\right)-\Phi\left(\frac{b_0-\theta}{\sigma_0}\right)\right]\right. \\
&+ \left.\frac{\sigma_0(\theta-b_0)}{2\pi}\exp\left\{\frac{-(b_0+\theta)^2}{2\sigma_0^2}\right\}-\frac{\sigma_0(\theta+b_0)}{2\pi}\exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}\right\} \\
&+ \frac{c_1}{1+c_1}\left\{(\theta^2+\sigma_0^2)\left[\Phi\left(\frac{-b_1+\theta}{\sigma_0}\right)-\Phi\left(\frac{-b_1-\theta}{\sigma_0}\right)\right]\right. \\
&+ \left.\frac{\sigma_0(\theta+b_1)}{2\pi}\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\}+\frac{\sigma_0(b_1-\theta)}{2\pi}\exp\left\{\frac{-(b_1+\theta)^2}{2\sigma_0^2}\right\}\right\} \\
&+ (\theta^2+\sigma_0^2)\left[\Phi\left(\frac{b_1-\theta}{\sigma_0}\right)-\Phi\left(\frac{b_0-\theta}{\sigma_0}\right)\right] \\
&+ \frac{\sigma_0}{2\pi}\left[(b_0+\theta)\exp\left\{\frac{-(b_0-\theta)^2}{2\sigma_0^2}\right\}-(b_1+\theta)\exp\left\{\frac{-(b_1-\theta)^2}{2\sigma_0^2}\right\}\right]
\end{aligned}
$$

3.7.D Performance Computations for Several Noise Distributions

Below are plots of D versus $\log T$ for several noise distributions and quickest detection procedures. The plots were obtained via the Markov approximation technique described in Chapter 2. Estimates of η are obtained by measuring the slope of the performance curves for large T and taking the inverse. The particular detectors that were used are indicated on the graphs.

All values of $\tilde{\eta}$ and η agree within 2%, and most are identical to an accuracy of three decimal places. Therefore, the approximation $\eta \approx \tilde{\eta}$ is valid.

Table D.1: Gaussian noise

	SNR = 0 dB		SNR = -10 dB	
test	*computed*	*measured*	*computed*	*measured*
linear	2.000	1.999	0.200	0.199
sign	1.139	1.139	0.126	0.126
dead-zone	1.493	1.493	0.161	0.161
robust, $\varepsilon = 0.1$	1.520	1.522	0.155	0.153
robust, $\varepsilon = 0.01$	1.822	1.812	0.193	0.192
robust, $\delta = 0.2$	1.495	1.495	0.136	0.137
robust, $\delta = 0.05$	1.723	1.721	0.179	0.177

Table D.2: Gauss-Gauss noise, $\varepsilon = 0.01$

	SNR = 0 dB		SNR = -10 dB	
test	*computed*	*measured*	*computed*	*measured*
linear	0.389	0.393	0.093	0.094
sign	2.006	2.006	0.243	0.243
dead-zone	2.603	2.605	0.309	0.309
robust, $\varepsilon = 0.01$	2.995	3.018	0.369	0.366

Table D.3: Gauss-Gauss noise, $\varepsilon = 0.1$

	SNR = 0 dB		SNR = -10 dB	
test	*computed*	*measured*	*computed*	*measured*
linear	0.730	0.732	0.138	0.139
sign	3.001	3.001	0.976	0.976
dead-zone	3.057	3.057	1.135	1.135
robust, $\varepsilon = 0.1$	3.040	3.060	1.198	1.191

Table D.4: Least favorable ε-contaminated noise, $\varepsilon = 0.01$

	SNR = 0 dB		SNR = -10 dB	
test	*computed*	*measured*	*computed*	*measured*
linear	1.511	1.515	0.188	0.191
sign	1.240	1.240	0.142	0.142
dead-zone	1.568	1.568	0.176	0.176
robust, $\varepsilon = 0.01$	1.780	1.780	0.204	0.203

Table D.5: Least favorable ε-contaminated noise, $\varepsilon = 0.1$

	SNR = 0 dB		SNR = -10 dB	
test	*computed*	*measured*	*computed*	*measured*
linear	1.285	1.298	0.189	0.190
sign	1.537	1.537	0.202	0.202
dead-zone	1.526	1.526	0.203	0.203
robust, $\varepsilon = 0.1$	1.579	1.567	0.233	0.233

Table D.6: Least favorable total variation noise, $\delta = 0.05$

	SNR = 0 dB		SNR = -10 dB	
test	*computed*	*measured*	*computed*	*measured*
linear	1.408	1.407	0.192	0.191
sign	1.353	1.353	0.162	0.162
dead-zone	1.615	1.615	0.190	0.190
robust, $\delta = 0.05$	1.742	1.741	0.214	0.213

Table D.7: Least favorable total variation noise, $\delta = 0.2$

	SNR = 0 dB		SNR = -10 dB	
test	*computed*	*measured*	*computed*	*measured*
linear	1.286	1.285	0.189	0.188
sign	1.538	1.538	0.211	0.211
dead-zone	1.522	1.522	0.198	0.198
robust, $\delta = 0.2$	1.580	1.583	0.241	0.240

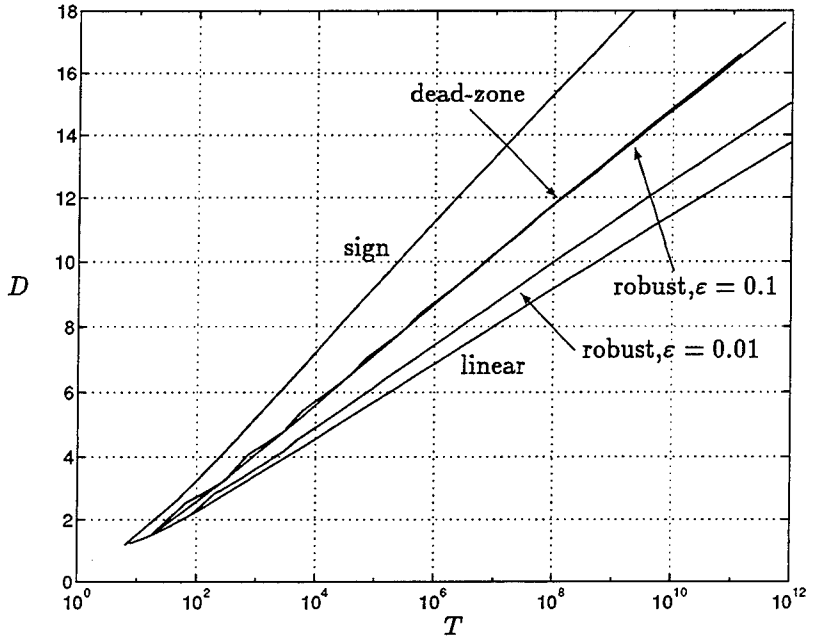

Figure 3.33: Performance for Gaussian noise: $\Psi = 0$ dB.

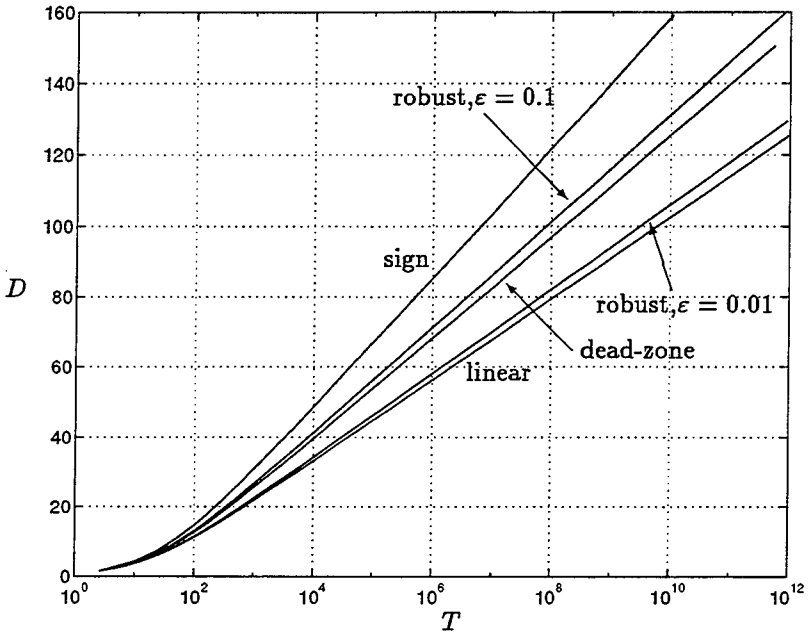

Figure 3.34: Performance for Gaussian noise: $\Psi = -10$ dB.

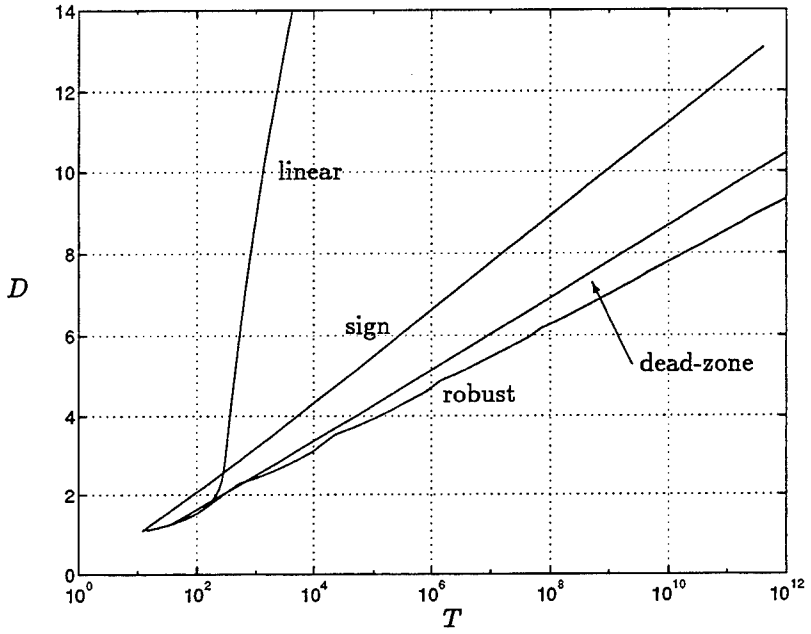

Figure 3.35: Performance for Gauss-Gauss noise: $\varepsilon = 0.01$, $\gamma = 100$, $\Psi = 0$ dB.

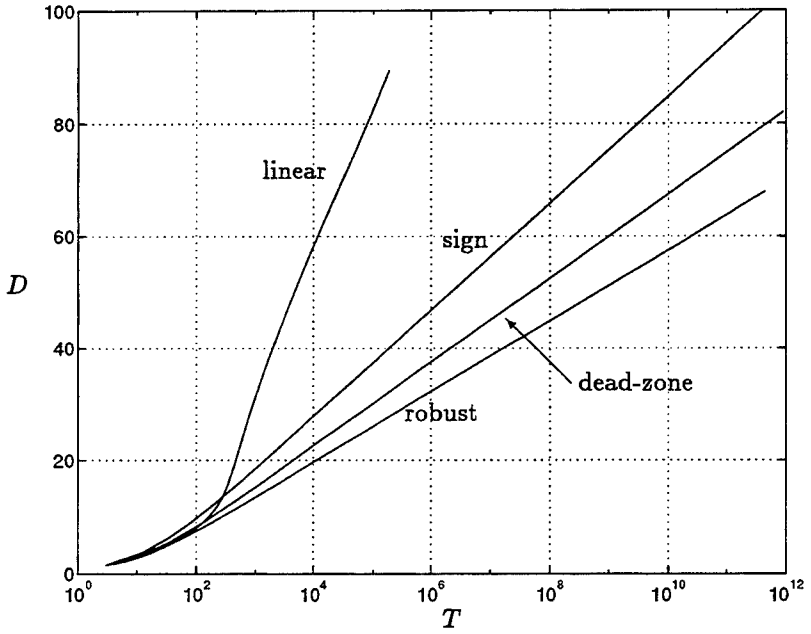

Figure 3.36: Performance for Gauss-Gauss noise: $\varepsilon = 0.01$, $\gamma = 100$, $\Psi = -10$ dB.

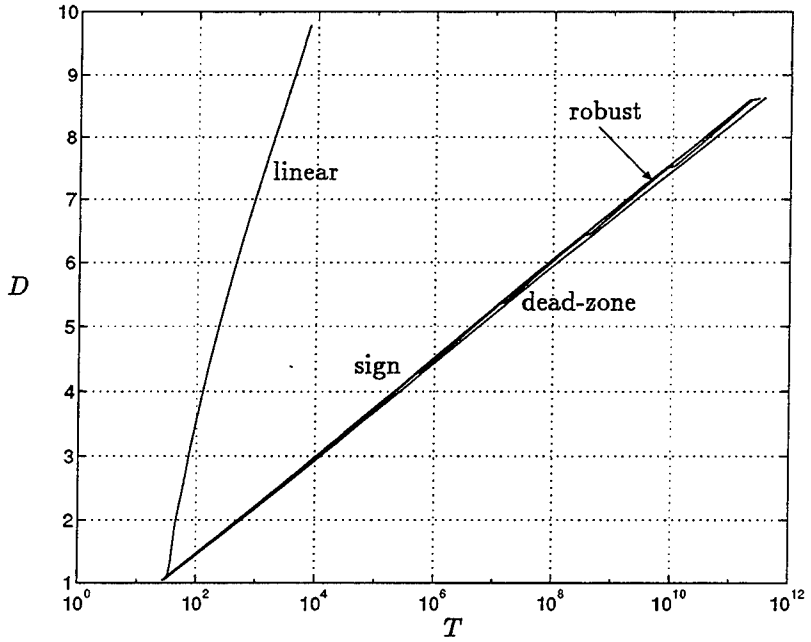

Figure 3.37: Performance for Gauss-Gauss noise: $\varepsilon = 0.1$, $\gamma = 100$, $\Psi = 0$ dB.

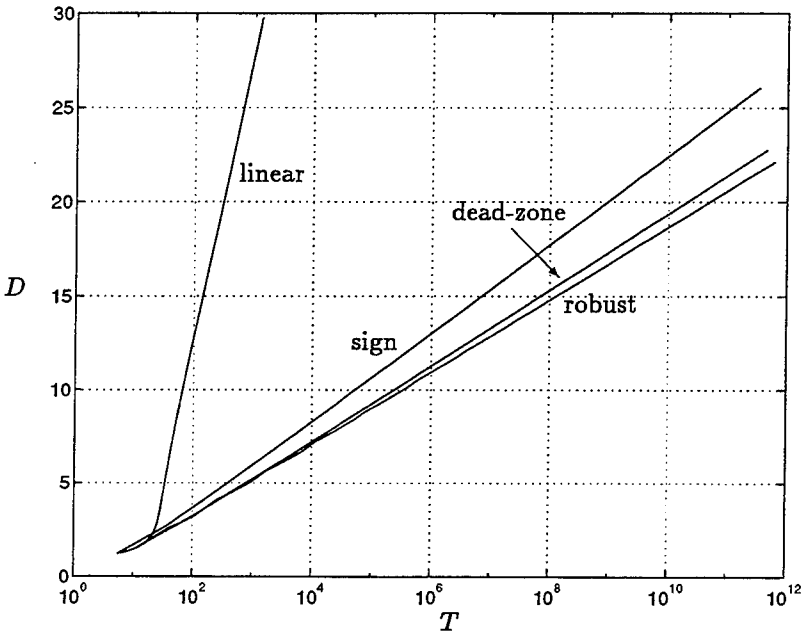

Figure 3.38: Performance for Gauss-Gauss noise: $\varepsilon = 0.1$, $\gamma = 100$, $\Psi = -10$ dB.

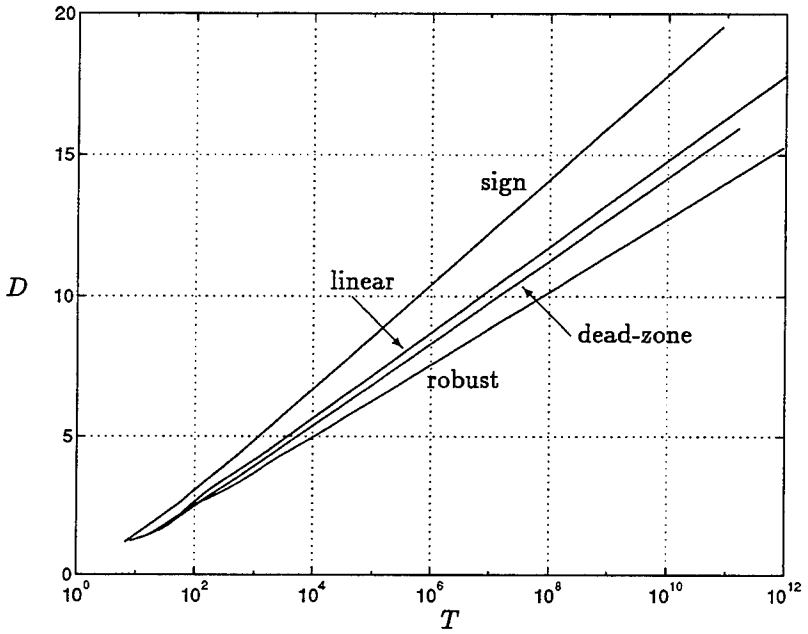

Figure 3.39: Performance for least favorable ε-contaminated noise: $\varepsilon = 0.01$, $\Psi = 0$ dB.

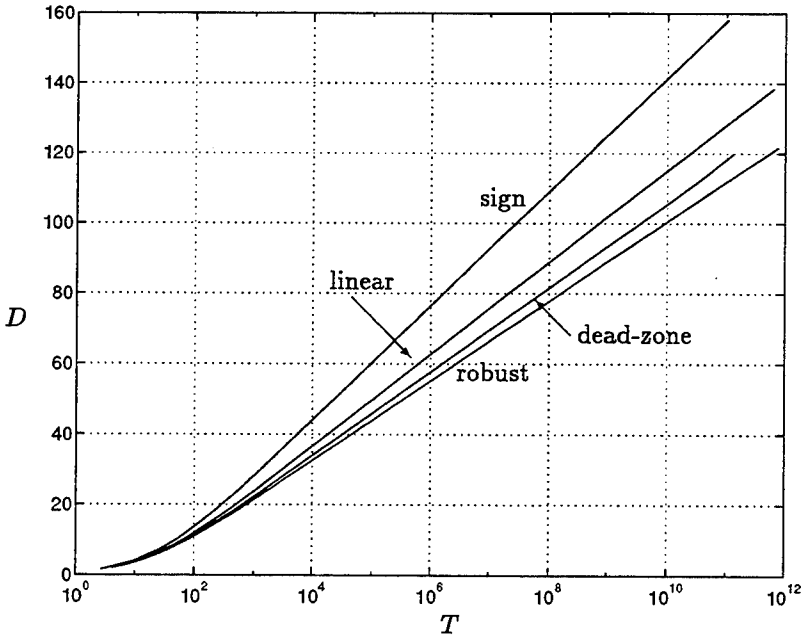

Figure 3.40: Performance for least favorable ε-contaminated noise: $\varepsilon = 0.01$, $\Psi = -10$ dB.

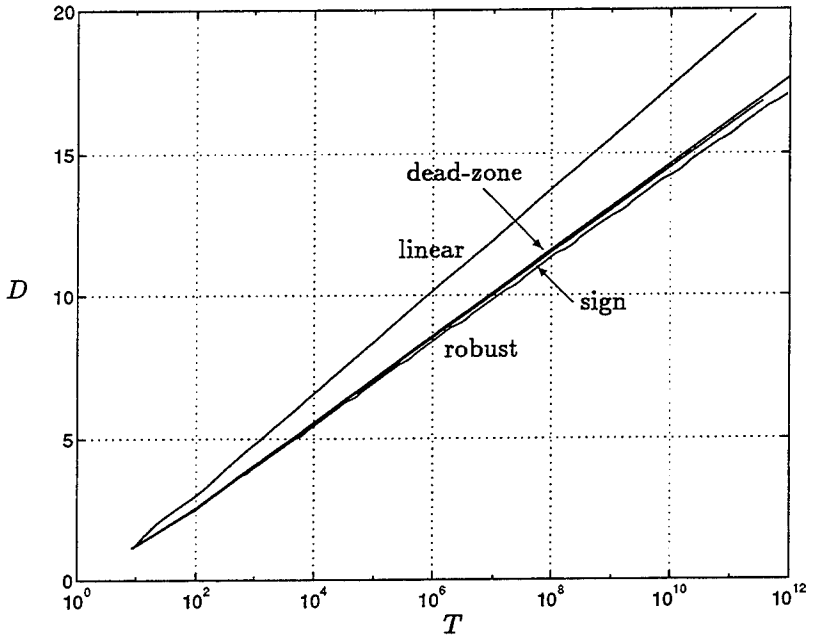

Figure 3.41: Performance for least favorable ε-contaminated noise: $\varepsilon = 0.1$, $\Psi = 0$ dB.

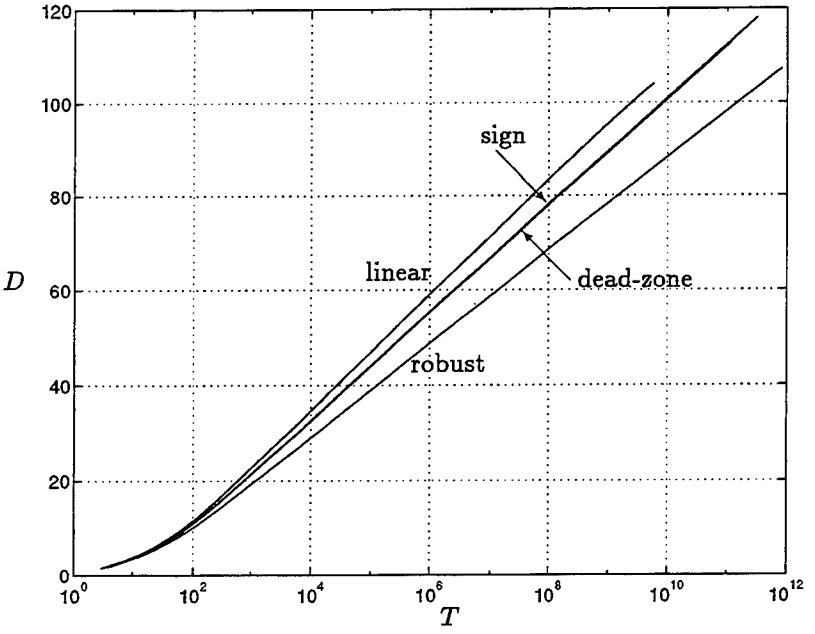

Figure 3.42: Performance for least favorable ε-contaminated noise: $\varepsilon = 0.1$, $\Psi = -10$ dB.

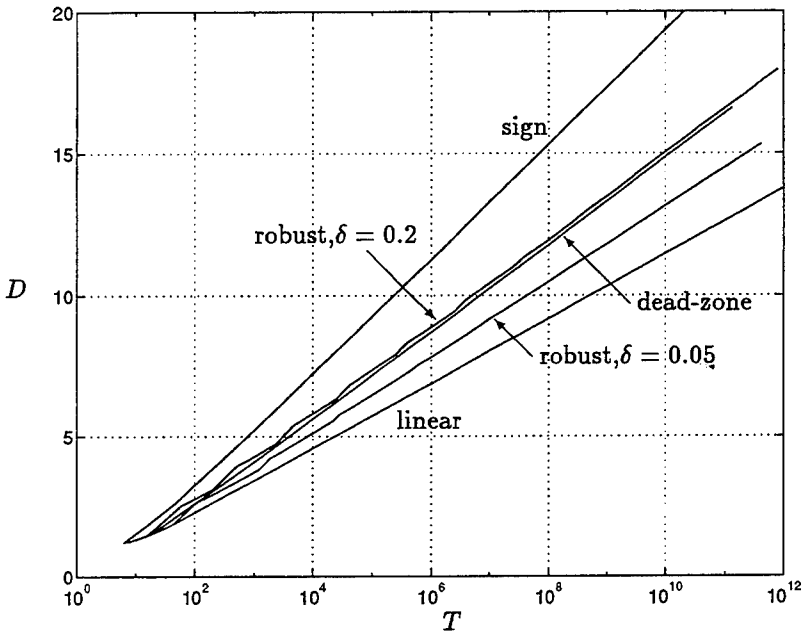

Figure 3.43: Performance for Gaussian noise: $\Psi = 0$ dB.

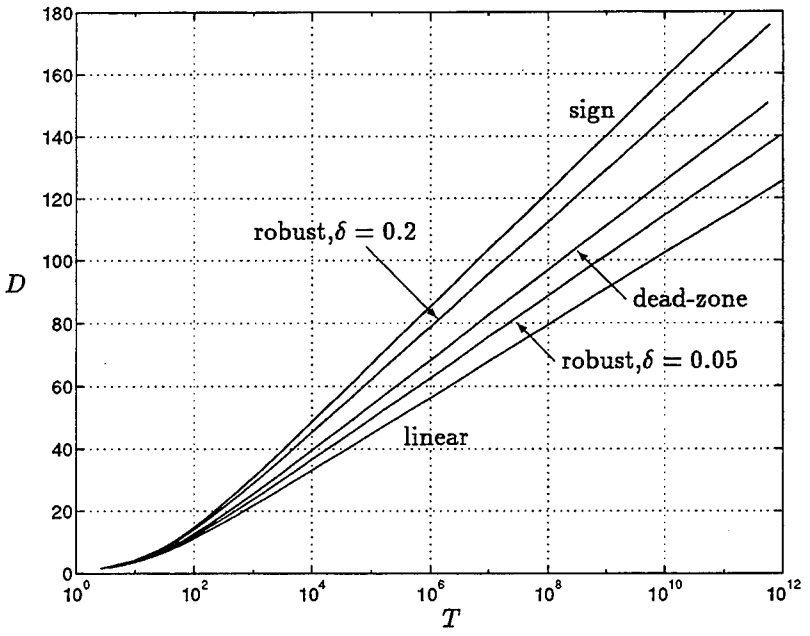

Figure 3.44: Performance for Gaussian noise: $\Psi = -10$ dB.

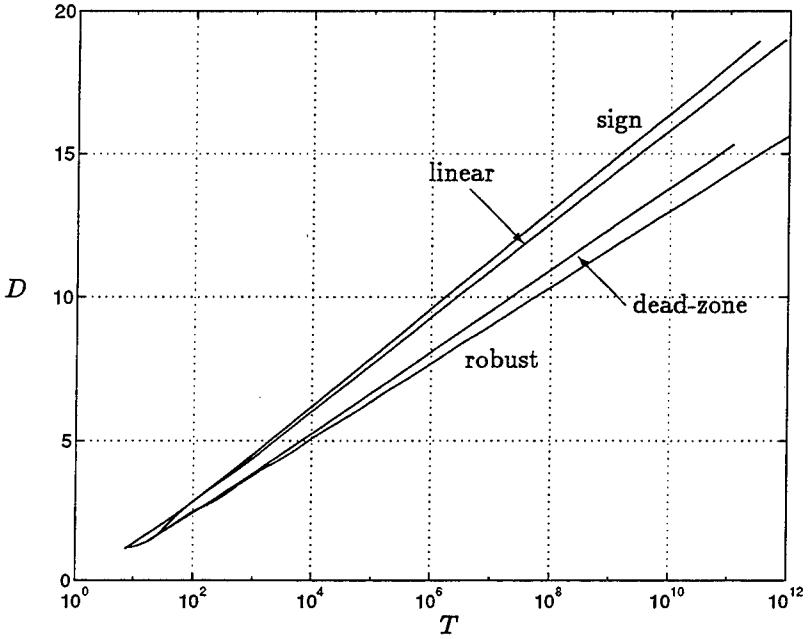

Figure 3.45: Performance for least favorable total variation noise: $\delta = 0.05$, $\Psi = 0$ dB.

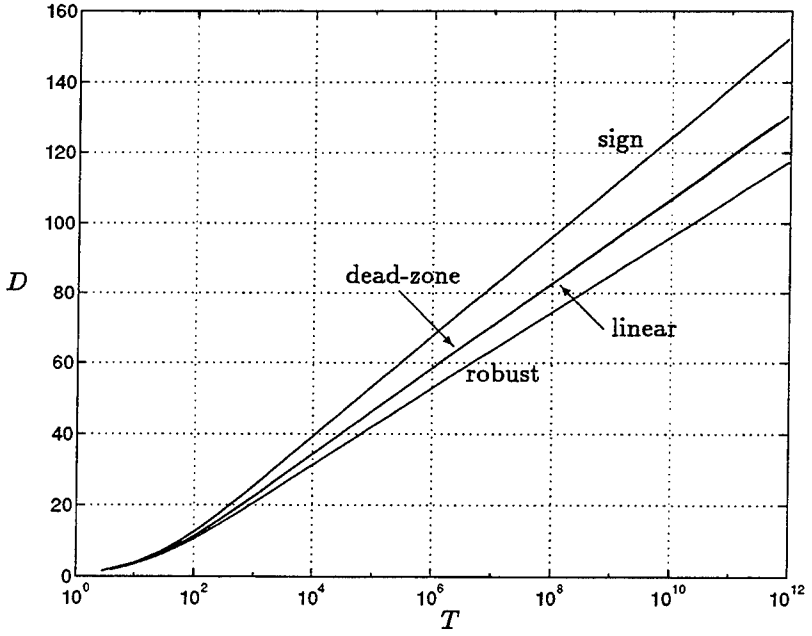

Figure 3.46: Performance for least favorable total variation noise: $\delta = 0.05$, $\Psi = -10$ dB.

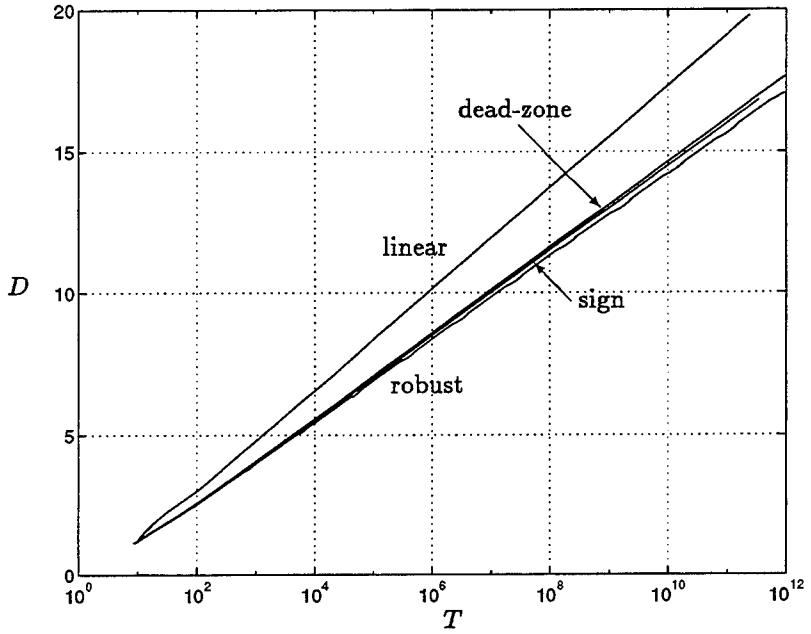

Figure 3.47: Performance for least favorable total variation noise: $\delta = 0.2$, $\Psi = 0$ dB.

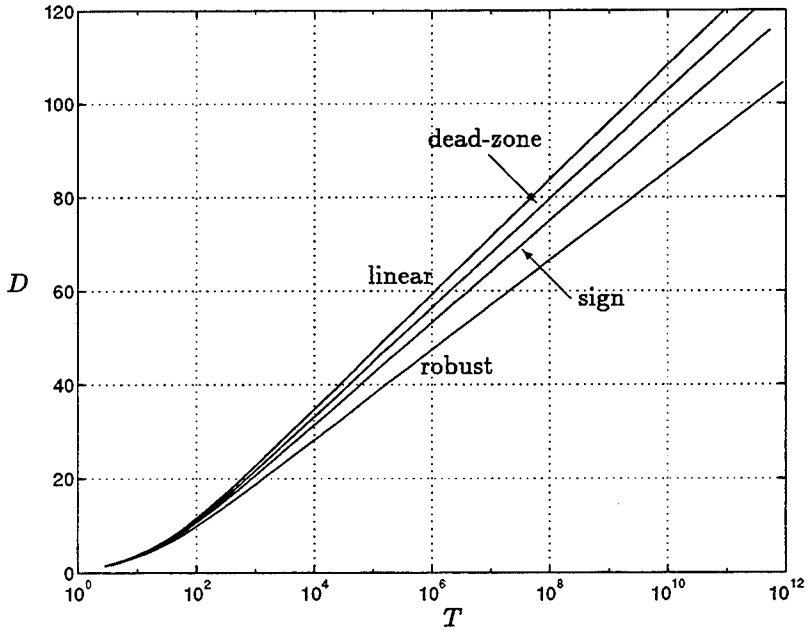

Figure 3.48: Performance for least favorable total variation noise: $\delta = 0.2$, $\Psi = -10$ dB.

3.7.E Efficacy Computations for the Weak Signal Case

The efficacy is

$$\mathcal{E} = \frac{\left(\int g'(x) f(x) dx\right)^2}{\int g^2(x) f(x) dx}$$

The computation of the two integrals, although tedious, is straightforward. Therefore, only the final expressions are given for each case.

Linear Detector, Gaussian Noise

$$\mathcal{E} = \frac{1}{\sigma_0^2}$$

Sign Detector, Gaussian Noise

$$\mathcal{E} = \frac{2}{\pi \sigma_0^2}$$

Dead-zone Limiter, Gaussian Noise

$$\mathcal{E} = \frac{1}{\pi \sigma_0^2} e^{-d^2/\sigma_0^2} \left[\Phi\left(\frac{-d}{\sigma_0}\right)\right]^{-1}$$

Robust Detector, Gaussian Noise

$$\mathcal{E} = \frac{\left\{\sigma_0^{-2}\left[2\Phi\left(k\sigma_0\right) - 1\right]\right\}^2}{2k^2\Phi\left(-k\sigma_0\right) - \sqrt{\frac{2}{\pi\sigma_0^2}} k e^{-k^2\sigma_0^2/2} + \sigma_0^{-2}\left[2\Phi\left(k\sigma_0\right) - 1\right]}$$

Linear Detector, Gauss-Gauss Noise

$$\mathcal{E} = \frac{1}{(1-\varepsilon)\sigma_0^2 + \varepsilon\sigma_1^2}$$

Sign Detector, Gauss-Gauss Noise

$$\mathcal{E} = \frac{2}{\pi}\left[\frac{1-\varepsilon}{\sigma_0} + \frac{\varepsilon}{\sigma_1}\right]$$

Dead-zone Limiter, Gauss-Gauss Noise

$$\mathcal{E} = \frac{\left\{\frac{1-\varepsilon}{\sigma_0}e^{-d^2/2\sigma_0^2} + \frac{\varepsilon}{\sigma_1}e^{-d^2/2\sigma_1^2}\right\}^2}{\pi\left[(1-\varepsilon)\Phi\left(\frac{-d}{\sigma_0}\right) + \varepsilon\Phi\left(\frac{-d}{\sigma_1}\right)\right]}$$

Robust Detector, Gauss-Gauss Noise

Let

$$I_1 = 2k^2\Phi\left(-k\sigma_0\right) - \sqrt{\frac{2}{\pi\sigma_0^2}}ke^{-k^2\sigma_0^2/2} + \sigma_0^{-2}\left[2\Phi\left(k\sigma_0\right) - 1\right]$$

and

$$I_2 = 2k^2\Phi\left(\frac{-k\sigma_0^2}{\sigma_1}\right) - \sqrt{\frac{2}{\pi}}\frac{\sigma_1}{\sigma_0^2}ke^{-k^2\sigma_0^4/2\sigma_1^2} + \frac{\sigma_1^2}{\sigma_0^4}\left[2\Phi\left(\frac{k\sigma_0^2}{\sigma_1}\right) - 1\right]$$

Then

$$\mathcal{E} = \frac{\left\{\sigma_0^{-2}\left[2(1-\varepsilon)\Phi\left(k\sigma_0\right) + 2\varepsilon\Phi\left(\frac{k\sigma_0^2}{\sigma_1}\right) - 1\right]\right\}^2}{(1-\varepsilon)I_1 + \varepsilon I_2}$$

Linear Detector, Least Favorable ε-contaminated Noise

$$\mathcal{E} = \left(\frac{2}{\pi\sigma_0^2}e^{-k^2\sigma_0^2/2}\left\{(1-\varepsilon)\left[\frac{2\sigma_0^2}{k} + \frac{2}{k^3}\right] - \varepsilon k\sigma_0^2\right\} + \sigma_0^2\left[2\Phi(k\sigma_0) - 1\right]\right)^{-1}$$

Sign Detector, Least Favorable ε-contaminated Noise

$$\mathcal{E} = \frac{2}{\pi\sigma_0^2}(1-\varepsilon)^2$$

Dead-zone Limiter, Least Favorable ε-contaminated Noise

$$\mathcal{E} = \begin{cases} \dfrac{(1-\varepsilon)e^{-d^2/\sigma_0^2}}{\sqrt{\frac{\pi}{2}}\sigma_0 ke^{-k^2\sigma_0^2/2} + \pi\sigma_0^2\left[\Phi\left(\frac{-d}{\sigma_0}\right) - \Phi(-k\sigma_0)\right]} & d \le k\sigma_0^2 \\ \sqrt{\frac{2}{\pi}}\frac{k(1-\varepsilon)}{\sigma_0}e^{k^2\sigma_0^2/2}e^{-kd} & d > k\sigma_0^2 \end{cases}$$

3.8 References

[1] R. N. Brandt and S. Karlin, "On the design and comparison of certain dichotomous experiments," Ann. Math. Stat., vol. 27, pp. 390-409, 1956.

[2] B. Broder, *Quickest Detection Procedures and Transient Signal Detection*, Ph.D. Thesis, Princeton University, 1990.

[3] W. A. Gardner, "A unifying view of second-order measures of quality for signal classification," *IEEE Trans. Commun.*, vol. COM-28, no. 6, June 1980.

[4] P. J. Huber, *Robust Statistics*, John Wiley & Sons, 1981.

[5] P. J. Huber, "A robust version of the probability ratio test," Ann. Math. Stat., vol. 36, pp. 1753-1758, 1965.

[6] S. A. Kassam and J. B. Thomas, "Dead-zone limiter: An application of conditional tests in nonparametric detection," *J. Acoust. Soc. Am.*, vol. 60, no. 4, pp. 857-862, October 1976.

[7] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897-1908, 1971.

[8] J. H. Miller, *Signal Detection in Non-Gaussian Noise*, Ph.D. Thesis, Princeton University, 1972.

[9] H. V. Poor, "On Robust Wiener Filtering," *IEEE Trans. Automat. Cont.*, vol. AC-25, no. 3, June 1980.

[10] J. B. Thomas, *An Introduction to Communication Theory and Systems*, Springer-Verlag, New York, 1988.

[11] H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Part I*, John Wiley and Sons, New York, 1968.

Chapter 4

Robust Quickest Detection Under Mean & Covariance Uncertainty

4.1 Introduction

In this chapter, we continue our investigation of robust quickest detection procedures. In the previous chapter, robust procedures were investigated for the case where the noise distributions were known only to lie in some uncertainty class. Here, we consider the case where the noise is multivariate Gaussian, and where the uncertainty exists in the mean vector and/or covariance matrix.

For the classical known-signal hypothesis testing problem in Gaussian noise, it is well known that the procedure that maximizes the probability of detection for a given false alarm probability (i.e., that satisfies the Neyman-Pearson criterion) is the *matched filter*, which is the filter that maximizes the output signal-to-noise ratio (SNR), followed by a comparator [9]. It is perhaps then not surprising that the matched filter is also the optimal processor for the quickest detection problem when the noise is Gaussian: the log-likelihood is the optimal processor in the sense of Lorden (see Chapter 2), and for the Gaussian case the log-likelihood function is exactly the matched filter.

The optimal processor when the noise is Gaussian is linear, and is based only on the first and second order statistics. If the noise is non-Gaussian, the optimum processor will in general be nonlinear, and hence the matched filter processor will be suboptimal. Unfortunately, the derivation of the optimum processor for non-Gaussian noise is not always straightforward, particularly if the noise cannot be characterized exactly. By comparison, the matched filter maximizing the SNR is a "common, simple and generally well-founded engineering technique [3]," and therefore may be an attractive option even when the noise is not Gaussian. [1]

Much work has been done in the area of robust matched filtering (for example, see [4] and references therein, and for a general treatment, see [8]); of particular interest in this chapter is the work on minimax robust discrete-time matched filtering of Verdu and Poor [10]. The main objective of this chapter is to formally establish the connection between robust matched filtering and robust quickest detection in multivariate Gaussian noise. Once this is done, we will see that all of the techniques from the former can be used directly to obtain solutions for the latter.

Section 2 contains the main result of this chapter. After stating the problem formally, it is shown that the minimax robust matched filter is exactly the optimal processor in the sense of Lorden. This is done by applying the minimax criterion directly to the asymptotic performance measure. Different types of signal and noise uncertainty are investigated in Section 3. It is shown that many of the results from [10] showing how to obtain the robust matched filters can also be used to derive the robust quickest detector. The asymptotic performance measures for the robust procedures are computed, and several examples are provided. In Section 4, two related issues are addressed. First, an alternative "minimax tuning" method which appears in [1] is shown to be equivalent to our approach. Finally, the computation of the asymptotic performance is discussed for the more general case when the noise is non-Gaussian.

[1] An exception is when the noise is impulsive. See [3] for details.

4.2 General Solution for the Robust Quickest Detector

Consider the following disorder problem. The multivariate real-valued independent random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are observed sequentially, where $\mathbf{x}_i$ is generated under H_0 for $i = 1, \ldots, m-1$ and under H_1 for $i = m, m+1, \ldots$. The two hypotheses are multivariate Gaussian:

$$H_0 \;:\; \mathbf{x}_i \sim \mathcal{N}(-\mathbf{s}, \Sigma)$$

$$H_1 \;:\; \mathbf{x}_i \sim \mathcal{N}(\mathbf{s}, \Sigma)$$

where $\mathbf{s} \in \Re^k$, $\Sigma \in \Re^{k \times k}$, and Σ is positive definite. The means are chosen to be symmetric for simplicity, but without loss of generality. Furthermore, it is known only that $(\mathbf{s}, \Sigma) \in \mathcal{S} \times \mathcal{N}$, where $\mathcal{S}$ and $\mathcal{N}$ are independent signal and noise uncertainty classes, respectively. [2] In this work we consider classes of the form:

$$\mathcal{S} \;=\; \{\mathbf{s} : \|\mathbf{s} - \mathbf{s}_0\| \leq \delta\}$$

$$\mathcal{N} \;=\; \{\Sigma : \|\Sigma - \Sigma_0\| \leq \varepsilon\}$$

Thus, the uncertainty is modelled as a deviation from the nominal parameters $\mathbf{s}_0$ and Σ_0. The particular norms used will be discussed in the next section. [3]

Page's test is defined here exactly as in Chapter 2; namely, the statistic

$$S_n = \max\{S_{n-1} + g(\mathbf{x}), 0\}$$

is recursively computed, and a disorder is declared when the stopping time

$$N = \inf\{n \mid S_n > h\}$$

[2] Notice that this is equivalent to saying that $(f_0, f_1) \in \mathcal{F}_0 \times \mathcal{F}_1$, where

$$\mathcal{F}_0 \;=\; \left\{f : f(\mathbf{x}) = \left|\tfrac{\Sigma}{2\pi}\right|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x}+\mathbf{s})^T\Sigma^{-1}(\mathbf{x}+\mathbf{s})\right\},\ \mathbf{s} \in \mathcal{S},\ \Sigma \in \mathcal{N}\right\}$$
$$\mathcal{F}_1 \;=\; \left\{f : f(\mathbf{x}) = \left|\tfrac{\Sigma}{2\pi}\right|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x}-\mathbf{s})^T\Sigma^{-1}(\mathbf{x}-\mathbf{s})\right\},\ \mathbf{s} \in \mathcal{S},\ \Sigma \in \mathcal{N}\right\}$$

In this chapter, the former notation will be used so that it is clear that the uncertainty lies only in the mean and covariance.

[3] It is not difficult to verify that $\mathcal{S}$ and $\mathcal{N}$ are also convex.

occurs, where $h > 0$ is some prespecified threshold. For Gaussian noise, the optimal processor is linear; that is, the processor $g(\mathbf{x}) = \mathbf{h}^T\mathbf{x}$, for some $\mathbf{h} \in \Re^k$. The SNR for a single snapshot is

$$\frac{|\langle \mathbf{h}, \mathbf{s} \rangle|^2}{\langle \mathbf{h}, \Sigma\mathbf{h} \rangle} \triangleq \rho(\mathbf{h}; \mathbf{s}, \Sigma)$$

and a direct application of the Cauchy-Schwarz inequality

$$|\langle \mathbf{h}, \mathbf{s} \rangle|^2 \leq \langle \mathbf{h}, \Sigma\mathbf{h} \rangle \langle \mathbf{s}, \Sigma^{-1}\mathbf{s} \rangle$$

reveals that ρ is maximized when $\mathbf{h}$ satisfies

$$\Sigma\mathbf{h} = \kappa\,\mathbf{s} \tag{4.1}$$

where κ is any nonzero real constant; in this case, $\rho(\mathbf{h}; \mathbf{s}, \Sigma) = \langle \mathbf{s}, \Sigma^{-1}\mathbf{s} \rangle$. [4] Equation (4.1) defines the well-known discrete-time matched filter, the processor that maximizes the SNR. Notice that $\rho(\mathbf{h}; \mathbf{s}, \Sigma)$ is the same under either H_0 or H_1, since we assumed that the respective mean vectors are symmetric.

The minimax solution for the robust matched filtering problem, $(\mathbf{h}_R, (\mathbf{s}_L, \Sigma_L))$, is the solution of

$$\max_{\mathbf{h} \in \mathcal{H}} \left\{ \min_{(\mathbf{s}, \Sigma) \in \mathcal{S} \times \mathcal{N}} \rho(\mathbf{h}; \mathbf{s}, \Sigma) \right\}$$

where (s_L, Σ_L) denotes the least favorable pair in $\mathcal{S} \times \mathcal{N}$, and $\Sigma_L\mathbf{h}_R = \kappa\,\mathbf{s}_L$. Here $\mathcal{H} = \Re^k$, but in general $\mathcal{H}$ can be an arbitrary Hilbert space. In [8], it is shown that a saddle point solution exists for this problem, that is:

$$\max_{\mathbf{h} \in \mathcal{H}} \rho(\mathbf{h}; \mathbf{s}_L, \Sigma_L) = \rho(\mathbf{h}_R; \mathbf{s}_L, \Sigma_L) = \min_{(\mathbf{s}, \Sigma) \in \mathcal{S} \times \mathcal{N}} \rho(\mathbf{h}_R; \mathbf{s}, \Sigma) \tag{4.2}$$

This allows the maximization and minimization to be determined separately rather than jointly. The following lemma (reworded slightly), which appears in [10], characterizes the saddle point solution:

[4] $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{y}$ is the *inner product* of $\mathbf{x}$ and $\mathbf{y}$.

Lemma 2: $(\mathbf{h}_R, (\mathbf{s}_L, \Sigma_L))$ *is a saddle point for the robust matched filtering problem if and only if*

1. $\Sigma_L \mathbf{h}_R = \mathbf{s}_L$,
2. $|\langle \mathbf{s}_L, \mathbf{h}_R \rangle| \leq |\langle \mathbf{s}, \mathbf{h}_R \rangle|, \ \forall \mathbf{s} \in \mathcal{S}$,
3. $0 \leq \langle \mathbf{h}_L, (\Sigma_L - \Sigma)\mathbf{h}_L \rangle, \ \forall \Sigma \in \mathcal{N}$

We now show that the robust processor for Page's test is $g_R(\mathbf{x}) = c\,\mathbf{h}_R^T\mathbf{x}$, where c is any positive real constant, thereby establishing a formal connection between minimax robust matched filtering and quickest detection.

Proposition 4: *If* $(\mathbf{h}_R, (\mathbf{s}_L, \Sigma_L))$ *is the minimax solution for the matched filtering problem, then* $(g_R, (\mathbf{s}_L, \Sigma_L))$ *is the asymptotic minimax solution for the quickest detection problem, where* $\Sigma_L \mathbf{h}_R = \kappa\,\mathbf{s}_L$, $\kappa \neq 0$, *and* $g_R(\mathbf{x}) = c\,\mathbf{h}_R^T\mathbf{x}$, $c > 0$.

Proof:

Let $f_{L0} \sim \mathcal{N}(-\mathbf{s}, \Sigma)$, and $f_{L1} \sim \mathcal{N}(\mathbf{s}, \Sigma)$, and assume $\kappa = 1$ for convenience (and without loss of generality, since κ can otherwise be incorporated into $\mathbf{s}_L$). We would like to show that (4.2) implies

$$\max_{g \in \mathcal{G}} \eta(g; \mathbf{s}_L, \Sigma_L) = \eta(g_R; \mathbf{s}_L, \Sigma_L) = \min_{(\mathbf{s}, \Sigma) \in \mathcal{S} \times \mathcal{N}} \eta(g_R; \mathbf{s}, \Sigma) \tag{4.3}$$

Since η is maximized when g is the log-likelihood ratio (cf. Chapter 2), the left equality is achieved when $g(\mathbf{x}) = \log \frac{f_{L1}(\mathbf{x})}{f_{L0}(\mathbf{x})} = 2\mathbf{s}_L^T \Sigma_L^{-1} \mathbf{x} = 2\mathbf{h}_R^T \mathbf{x}$. However, recall that *i.)* when $g(\mathbf{x})$ is the log-likelihood, $\eta = \tilde{\eta}$, and *ii.)* $\tilde{\eta}$ is invariant to scale changes. Thus, the left equality is also achieved for $g(\mathbf{x}) = c\,\mathbf{h}_R^T\mathbf{x}$ where c is any positive real constant.

Now we need to show that the right equality holds when $(\mathbf{s}_L, \Sigma_L)$ is the least favorable pair for the robust matched filtering problem. To do this, first recall the

definition of the lower bound $\widetilde{\eta}$ (cf. Chapter 2) for Page's test implemented with processor $g(\mathbf{x})$:

$$\widetilde{\eta} = \omega_0 \mathrm{E}\left[g(\mathbf{x}) \mid H_1\right] \tag{4.4}$$

where ω_0 is the non-zero root of the moment generating function equality

$$\mathrm{E}\left[\exp\{\omega_0 g(\mathbf{x})\} \mid H_0\right] = 1$$

Let $f_0 \sim \mathcal{N}(-\mathbf{s}, \Sigma)$ and $f_1 \sim \mathcal{N}(\mathbf{s}, \Sigma)$. When the robust processor is used, ω_0 satisfies

$$1 = \int e^{\omega_0 g_R(\mathbf{x})} f_0(\mathbf{x}) d\mathbf{x} = \int e^{\mathbf{v}^T \mathbf{x}} f_0(\mathbf{x}) d\mathbf{x} \tag{4.5}$$

where $\mathbf{v} \stackrel{\Delta}{=} c\omega_0 \mathbf{h}_R$. Now observe that this last expression is just the vector moment generating function as a function of $\mathbf{v}$. Therefore,

$$1 = \exp\{-\mathbf{v}^T\mathbf{s} + \frac{1}{2}\mathbf{v}^T\Sigma\mathbf{v}\}$$

Taking the log of both sides, rearranging terms, and substituting back in for $\mathbf{v}$, we have

$$c\omega_0 \mathbf{h}_R^T \mathbf{s} = \frac{1}{2} c^2 \omega_0^2 \mathbf{h}_R^T \Sigma_L \mathbf{h}_R$$

of which the nonzero solution is

$$\omega_0 = \frac{2\mathbf{h}_R^T\mathbf{s}}{c\,\mathbf{h}_R^T\Sigma_L\mathbf{h}_R}$$

Also,

$$\mathrm{E}\left[g_R(\mathbf{x}) \mid H_1\right] = c\,\mathbf{h}_R^T\mathbf{s}$$

Therefore, the lower bound (4.4) is

$$\widetilde{\eta}(g_R; \mathbf{s}, \Sigma) = 2\frac{\left|\mathbf{h}_R^T\mathbf{s}\right|^2}{\mathbf{h}_R^T\Sigma\mathbf{h}_R} = 2\frac{|\langle \mathbf{h}_R, \mathbf{s}\rangle|^2}{\langle \mathbf{h}_R, \Sigma\mathbf{h}_R\rangle} \tag{4.6}$$

(Notice that this is independent of the scale factor c, as expected).

The right expression in (4.3) can now be lower bounded as

$$\begin{aligned}\min_{(\mathbf{s},\Sigma)\in\mathcal{S}\times\mathcal{N}} \eta(g_R;\mathbf{s},\Sigma) &\geq \min_{(\mathbf{s},\Sigma)\in\mathcal{S}\times\mathcal{N}} 2\frac{|\langle \mathbf{h}_R,\mathbf{s}\rangle|^2}{\langle \mathbf{h}_R,\Sigma\mathbf{h}_R\rangle} \\ &= 2\min_{(\mathbf{s},\Sigma)\in\mathcal{S}\times\mathcal{N}} \rho(\mathbf{h}_R;\mathbf{s},\Sigma) \\ &= 2\rho(\mathbf{h}_R;\mathbf{s}_L,\Sigma_L) \end{aligned} \tag{4.7}$$

where the last expression follows from (4.2), and $(\mathbf{s}_L,\Sigma_L)$ is the least favorable pair for the robust matched filtering problem. Now

$$\begin{aligned}\rho(\mathbf{h}_R;\mathbf{s}_L,\Sigma_L) &= \langle \mathbf{s}_L,\Sigma_L^{-1}\mathbf{s}_L\rangle \\ &= \frac{1}{2}I(f_{1L},f_{0L}) \\ &= \frac{1}{2}\eta(g_R;\mathbf{s}_L,\Sigma_L) \end{aligned} \tag{4.8}$$

where $I(\cdot,\cdot)$ is the K-L divergence. Thus, (4.7) and (4.8) imply that

$$\min_{(\mathbf{s},\Sigma)\in\mathcal{S}\times\mathcal{N}} \eta(g_R;\mathbf{s},\Sigma) \geq \eta(g_R;\mathbf{s}_L,\Sigma_L) \tag{4.9}$$

Conversely, since the least-favorable pair lies in $\mathcal{S}\times\mathcal{N}$, we also have that

$$\begin{aligned}\min_{(\mathbf{s},\Sigma)\in\mathcal{S}\times\mathcal{N}} \eta(g_R;\mathbf{s},\Sigma) &= \min\left\{\min_{(\mathbf{s},\Sigma)\neq(\mathbf{s}_L,\Sigma_L)} \eta(g_R;\mathbf{s},\Sigma),\ \eta(g_R;\mathbf{s}_L,\Sigma_L)\right\} \\ &\leq \eta(g_R;\mathbf{s}_L,\Sigma_L) \end{aligned} \tag{4.10}$$

Finally, (4.9) and (4.10) together imply

$$\min_{(\mathbf{s},\Sigma)\in\mathcal{S}\times\mathcal{N}} \eta(g_R;\mathbf{s},\Sigma) = \eta(g_R;\mathbf{s}_L,\Sigma_L)$$

and so (4.3) holds. ∎

We have seen that for the Gaussian case, the least favorable pairs can be obtained via the matched filtering formulation. Now, we would like to state some of the results for particular types of signal and noise uncertainty classes.

4.3 Particular Solutions for Various Uncertainty Classes

4.3.1 Signal Uncertainty

In this section, three signal uncertainty classes are considered, all of which are based on the l_p norm; it is assumed that the noise covariance, Σ_0, is fixed and known. In [10], necessary and sufficient conditions for $(\mathbf{h}_L, (\mathbf{s}_L, \Sigma_0))$ to be a saddle point solution of the minimax robust matched filtering problem (i.e., to satisfy (4.2)) are given. Proposition 1 of the previous section confirms that these same conditions can be used to obtain the optimal processor for the quickest detection problem. The asymptotic performance measures for the nominal and robust procedures are determined below. Finally, a simple example illustrates the utility of the previous results from robust matched filtering in the design of robust quickest detectors.

Types of Signal Uncertainty

Since the robust processor is that filter which is matched to the least favorable signal, we have that $\mathbf{h}_L = \Sigma_0^{-1}\mathbf{s}_L$ regardless of the choice of the class $\mathcal{S}$. The three types of uncertainty classes are listed below, along with the necessary and sufficient conditions derived in [10]. [5]

- *Mean-absolute distortion (l_1 norm):* $\mathcal{S}_1 = \{\| \mathbf{s} - \mathbf{s}_0 \|_1 \leq \Delta\}$

  For $i = 0, 1, \ldots, k-1$:

  $$s_{Li} = s_{0i} - \delta_i^2 \mathrm{sgn}(h_{Li})$$

  where

  $$\delta_i = 0 \text{ if } |h_{Li}| < M \triangleq \max_{j=0,\ldots,k-1} |h_{Lj}|$$

[5] For the most part, the notation we use here is the same as in [10].

and

$$\sum_{i=0}^{k-1} \delta_i^2 = \Delta$$

is satisfied. If the noise is uncorrelated, that is $\Sigma_0 = \text{diag}(\sigma_0^2, \ldots, \sigma_{k-1}^2)$, then the following closed-form expression can be derived:

$$h_{Li} = \begin{cases} h_{0i}, & \mid h_{0i} \mid \leq C \\ C\text{sgn}(h_{0i}), & \mid h_{0i} \mid > C \end{cases}$$

where C satisfies

$$\sum_{i=0}^{k-1} \sigma_i^2 (\mid h_{0i} \mid - C)^+ = \Delta$$

- *Mean-square distortion (l_2 norm):* $\mathcal{S}_2 = \{\| \mathbf{s} - \mathbf{s}_0 \|_2 \leq \Delta\}$

$$\mathbf{h}_L = (\Sigma_0 + \sigma_s^2 I)^{-1} \mathbf{s}_0$$

$$\mathbf{s}_L = \mathbf{s}_0 - \sigma_s^2 \mathbf{h}_L$$

where σ_s^2 is obtained by solving

$$\Delta = \sigma_s^2 \| (\Sigma_0 + \sigma_s^2 I)^{-1} \mathbf{s}_0 \|$$

- *Maximum-absolute distortion (l_∞ norm):* $\mathcal{S}_3 = \{\| \mathbf{s} - \mathbf{s}_0 \|_\infty \leq \Delta\}$

For $i = 0, 1, \ldots, k-1$:

$$s_{Li} = \begin{cases} s_{0i} - \Delta, & h_{Li} > 0 \\ s_{0i} + \Delta, & h_{Li} < 0 \end{cases}$$

If the noise is uncorrelated, then

$$s_{Li} = \begin{cases} s_{0i} - \Delta, & \Delta < s_{0i} \\ 0, & -\Delta \leq s_{0i} \leq \Delta \\ s_{0i} + \Delta, & s_{0i} < -\Delta \end{cases}$$

Unfortunately, a general closed-form solution is only available for the class $\mathcal{S}_2$; for the others, a closed-form solution is available only for uncorrelated noise. It has been suggested that numerical techniques might be used to obtain $\mathbf{h}_L$ for $\mathcal{S}_1$ and $\mathcal{S}_3$. This point is discussed further at the end of this section.

Asymptotic Performance Under Signal Uncertainty

We would like to compare the robust quickest detector to the nominal version (that is, the one which is optimal for the nominal parameters $(\mathbf{s}_0, \Sigma_0)$) for different values of $\mathbf{s}$. To do this, we compute $\tilde{\eta}$ for each case of interest, as shown below.

In the proof of Proposition 1, the lower bound on asymptotic performance for the robust procedure ($g_R(\mathbf{x}) = \mathbf{h}_R^T\mathbf{x}$) when the true parameters are $(\mathbf{s}, \Sigma)$ was derived in the proof of Proposition 1; the result is given in (4.6). This same derivation applies to the case where the assumed operating point is $(\mathbf{s}_1, \Sigma_0)$, but the true pair is $(\mathbf{s}_2, \Sigma_0)$; one simply makes the substitutions $g_R(\mathbf{x}) \leftarrow g_1(\mathbf{s}) \triangleq \mathbf{h}_1^T\mathbf{x}$, $\mathbf{h}_R \leftarrow \mathbf{h}_1 \triangleq \Sigma_0^{-1}\mathbf{s}_1$ and $\mathbf{s} \leftarrow \mathbf{s}_2$ in (4.6). The result is

$$\tilde{\eta}(g_1; \mathbf{s}_2, \Sigma_0) = 2\frac{|\langle \mathbf{h}_1, \mathbf{s}_2 \rangle|^2}{\langle \mathbf{h}_1, \Sigma_0 \mathbf{h}_1 \rangle} \tag{4.11}$$

Thus, $\tilde{\eta}$ for the robust test operating at the nominal point is given by simply substituting $\mathbf{s}_L \rightarrow \mathbf{s}_1$ and $\mathbf{s}_0 \rightarrow \mathbf{s}_2$. This gives:

$$\tilde{\eta}(g_R; \mathbf{s}_0, \Sigma_0) = 2\frac{\left|\mathbf{s}_L^T\Sigma_0^{-1}\mathbf{s}\right|^2}{\mathbf{s}_L^T\Sigma_0^{-1}\mathbf{s}_L}$$

Similarly, $\tilde{\eta}$ for the nominal procedure when the least favorable signal is present is obtained when $g_0(\mathbf{x}) = \mathbf{h}_0^T\mathbf{x}$, where $\mathbf{s}_0 \rightarrow \mathbf{s}_1$ and $\mathbf{s}_L \rightarrow \mathbf{s}_2$ in (4.11); we have:

$$\tilde{\eta}(g_0; \mathbf{s}_L, \Sigma_0) = 2\frac{\left|\mathbf{s}_0^T\Sigma_0^{-1}\mathbf{s}_L\right|^2}{\mathbf{s}_0^T\Sigma_0^{-1}\mathbf{s}_0}$$

In Chapter 2, we saw that when the log-likelihood ratio is used to process the data, then $\eta = \tilde{\eta}$, and $\tilde{\eta}$ is equal to the Kullback-Leibler (K-L) divergence. When

$f_0 \sim \mathcal{N}(-\mathbf{s}, \Sigma)$ and $f_1 \sim \mathcal{N}(\mathbf{s}, \Sigma)$, it is easy to verify that

$$I(f_1, f_0) = 2\mathbf{s}^T \Sigma^{-1} \mathbf{s}$$

Thus, $\tilde{\eta}(g_0; \mathbf{s}_0, \Sigma_0) = 2\mathbf{s}_0^T \Sigma_0^{-1} \mathbf{s}_0$ and $\tilde{\eta}(g_L; \mathbf{s}_L, \Sigma_0) = 2\mathbf{s}_L^T \Sigma_0^{-1} \mathbf{s}_L$.

Example

Here, we present an example which graphically illustrates the decision regions for each type of signal uncertainty, following the general design procedures detailed earlier. It is desired to design a robust quickest detection procedure where the parameters are nominally

$$\mathbf{s}_0 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \quad \text{and} \quad \Sigma_0 = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$$

with an uncertainty parameter of $\Delta = \frac{1}{2}$.

- $\underline{\mathbf{s} \in \mathcal{S}_1}$

  $$\mathbf{h}_0 = \Sigma_0^{-1} \mathbf{s}_0 = \begin{pmatrix} 0.3 \\ 1 \end{pmatrix}$$

  Choose C such that

  $$10\,(0.3 - C)^+ + (1 - C)^+ = \Delta$$

  When $\Delta = \frac{1}{2}$, $C = \frac{1}{2}$. Thus,

  $$\mathbf{h}_L = \begin{pmatrix} 0.3 \\ 0.5 \end{pmatrix} \quad \text{and} \quad \mathbf{s}_L = \Sigma_0 \mathbf{h}_L = \begin{pmatrix} 3 \\ 0.5 \end{pmatrix}$$

  We can now compare the asymptotic performance for the nominal and robust procedures when the observations are $\mathcal{N}(\pm\mathbf{s}_L, \Sigma_0)$. Using the results shown earlier

in this section, we have

$$\widetilde{\eta}(g_0; \mathbf{s}_L, \Sigma_0) = 2 \times \frac{\left| (3 \quad 1) \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 0.5 \end{pmatrix} \right|^2}{(3 \quad 1) \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 1 \end{pmatrix}} = 2.063$$

and

$$\widetilde{\eta}(g_L; \mathbf{s}_L, \Sigma_0) = 2 \times (3 \quad 0.5) \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 0.5 \end{pmatrix} = 2.300$$

Therefore, for the least favorable signal $\mathbf{s}_L$, the robust quickest detector outperforms the nominal procedure, as expected.

- $\underline{\mathbf{s} \in \mathcal{S}_2}$

  First, solve for σ_s^2, where

$$\sigma_s^2 \left\| (\Sigma_0 + \sigma_s^2 I)^{-1} \mathbf{s}_0 \right\|_2 = \sigma_s^2 \left\| \begin{pmatrix} \frac{3}{10+\sigma_s^2} \\ \frac{1}{1+\sigma_s^2} \end{pmatrix} \right\|_2 = \Delta$$

  As pointed out in [10], the left side is monotone increasing in σ_s^2. Therefore, it is easy to iteratively determine σ_s^2 (for example, via a bisection routine): the result for $\Delta = \frac{1}{2}$ is $\sigma_s^2 = 0.8079$. Thus

$$\mathbf{h}_L = \begin{pmatrix} 0.278 \\ 0.553 \end{pmatrix} \quad \text{and} \quad \mathbf{s}_L = \begin{pmatrix} 2.775 \\ 0.553 \end{pmatrix}$$

  For this signal uncertainty class, $\widetilde{\eta}(g_0; \mathbf{s}_L, \Sigma_0) = 2.021$ and $\widetilde{\eta}(g_L; \mathbf{s}_L, \Sigma_0) = 2.152$.

- $\underline{\mathbf{s} \in \mathcal{S}_3}$

  We have directly that

$$\mathbf{s}_L = \begin{pmatrix} 3 - \Delta \\ 1 - \Delta \end{pmatrix} = \begin{pmatrix} 2.5 \\ 0.5 \end{pmatrix}$$

and so

$$\mathbf{h}_L = \Sigma_0^{-1}\mathbf{s}_L = \begin{pmatrix} 0.25 \\ 0.5 \end{pmatrix}$$

Finally, $\tilde{\eta}(g_0; \mathbf{s}_L, \Sigma_0) = 1.645$ and $\tilde{\eta}(g_L; \mathbf{s}_L, \Sigma_0) = 1.750$.

Notice that for fixed Δ, both $\tilde{\eta}(g_0; \mathbf{s}_L, \Sigma_0)$ and $\tilde{\eta}(g_L; \mathbf{s}_L, \Sigma_0)$ decrease with each class. This fact is not surprising since $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \mathcal{S}_3$. We see that the asymptotic performance is inversely proportional to the amount of assumed uncertainty, a fact that agrees with intuition.

The uncertainty regions for the above examples are graphically illustrated in Figures 4.1 through 4.3. Let $g(\mathbf{x}) = \mathbf{h}^T\mathbf{x}$, where $\mathbf{h}$ is an arbitrary linear processor. Since $\mathrm{E}[g(\mathbf{x}) \mid H_1] = -\mathrm{E}[g(\mathbf{x}) \mid H_0] > 0$, observe that the decision regions in each case are separated by the hyperplane $\mathbf{h}^T\mathbf{x} = 0$ (the hyperplane is simply a line in these examples). For each signal class, the separating hyperplanes for the nominal and robust procedures are shown. The slopes of the boundaries are dependent on the noise covariance Σ_0. For example, when $\Sigma_0 = I$, the hyperplanes are perpendicular to the line connecting $-\mathbf{s}_0$ and $\mathbf{s}_0$ (or $-\mathbf{s}_L$ and $\mathbf{s}_L$); on the other hand, when the condition number of the covariance is not unity (the eigenvalues are not all the same), as in our example, the hyperplanes will be "skewed" with respect to this line.

We now examine the design tradeoffs involved in choosing Δ. In Figure 4.4, the asymptotic performances of the nominal and robust procedures are compared as a function of Δ for the class $\mathcal{S}_2$ (similar plots can be obtained for the other two classes). The nominal pair $(\mathbf{s}_0, \Sigma_0)$ is the same as in the previous example. Each plot of $\tilde{\eta}$ versus Δ is labelled with a pair of the form $(g, \mathbf{s})$, which indicates that the procedure with processor $g(\mathbf{x})$ was used, but that the true mean vector was $\mathbf{s}$ (there is still no uncertainty in the covariance matrix). Notice that the robust procedure implemented when the true signal is the least favorable, $\mathbf{s}_L$, outperforms the nominal procedure for all choices of Δ. However, this is at the price of reduced performance

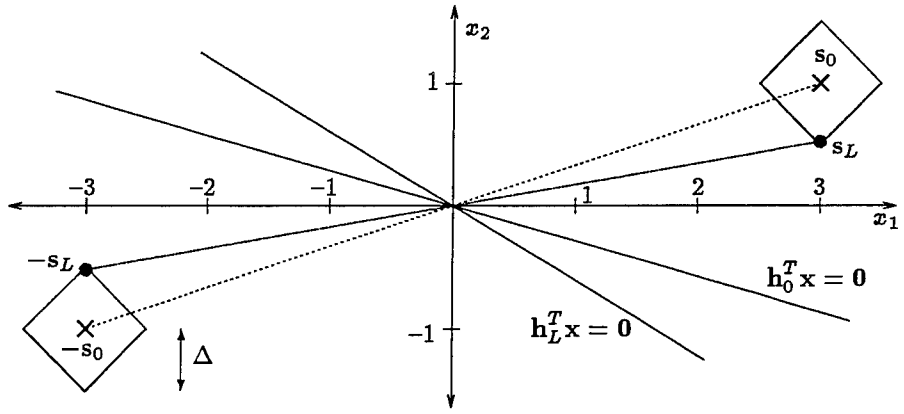

Figure 4.1: Signal uncertainty class $\mathcal{S}_1$.

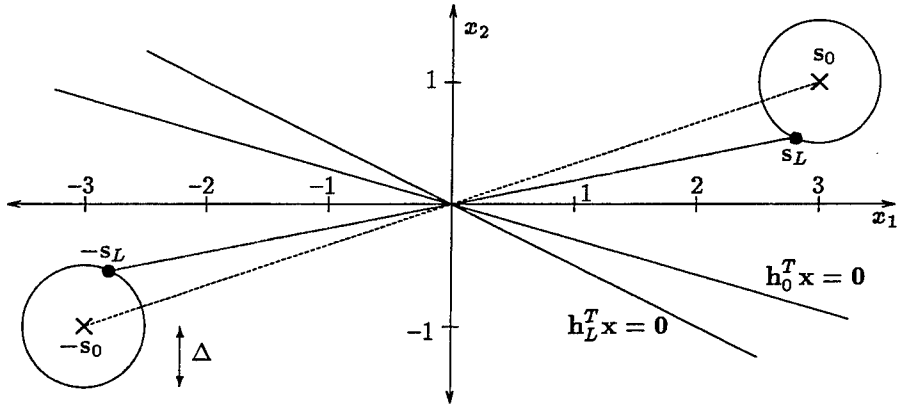

Figure 4.2: Signal uncertainty class $\mathcal{S}_2$.

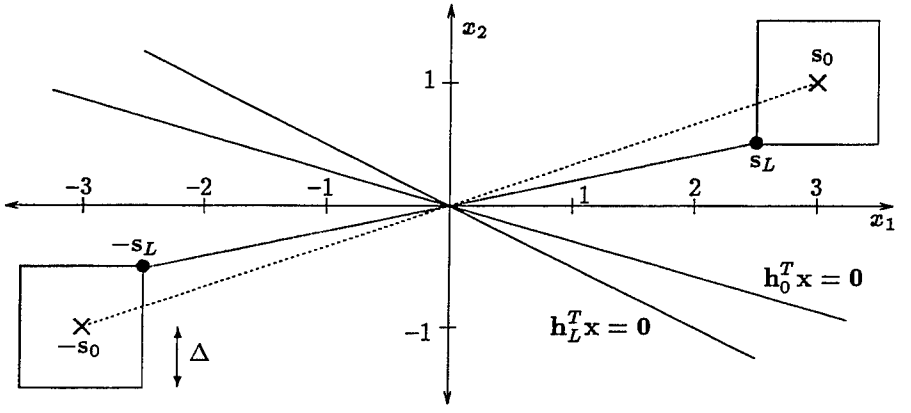

Figure 4.3: Signal uncertainty class $\mathcal{S}_3$.

when the signal is in fact $\mathbf{s}_0$.

Now suppose that a robust quickest detection procedure is designed for a signal distortion of Δ', but that the true distortion is Δ; this scenario is depicted in Figure 4.5. Observe that the procedure with $\Delta' = 1.0$ exhibits performance that is almost identical to the optimal procedure (designed for distortion Δ) when the $0.75 \leq \Delta \leq 1.5$, but that the performance degenerates with respect to the optimal and nominal procedures for $\Delta \leq 0.5$. Similarly, the performance of the procedure designed with $\Delta' = 0.25$ is reasonable for $\Delta \leq 0.5$, but declines compared to the optimal procedure over the rest of the interval. This illustrates a robustness of a different sort: namely, that good performance can be obtained using the robust procedure even if there is some mismatch in the assumed and actual levels of distortion.

In Chapter 3, the performance of two procedures were compared by computing the "robustness index," which was the ratio of the $\widetilde{\eta}$'s for each. This approach is also used in [10] in the context of minimax robust matched filtering in discrete time, and, while not included here, an analogous analysis can be used to compare robust quickest detectors.

Obtaining $\mathbf{s}_L$ and $\mathbf{h}_L$ When the Noise is Correlated

For classes $\mathcal{S}_1$ and $\mathcal{S}_3$ when the noise is correlated (Σ_0 is not diagonal), the least favorable signals can only be written as a function of the robust filter; that is, $\mathbf{s}_L = \mathbf{s}_L(\mathbf{h}_L)$. In these cases, $\mathbf{s}_L$ and $\mathbf{h}_L$ must be determined numerically. One possible approach to accomplish this is outlined below.

For both $\mathcal{S}_1$ and $\mathcal{S}_3$, the uncertainty region is bounded by the hyperplanes described by the equation $\|\mathbf{s} - \mathbf{s}_0\| = \Delta$. For example, when $\mathbf{s}_0 = (3 \;\; 1)^T$ and the l_∞ norm is used, the region is just a square, as shown in Figure 4.3. In Figure 4.6, a blowup of $\mathcal{S}_3$ is shown for $\Delta = \frac{1}{2}$. Notice that the region is simply the intersection of the four half-planes (which are just lines in this case) $x_1 \geq 2.5$, $x_1 \leq 3.5$, $x_2 \geq 0.5$,

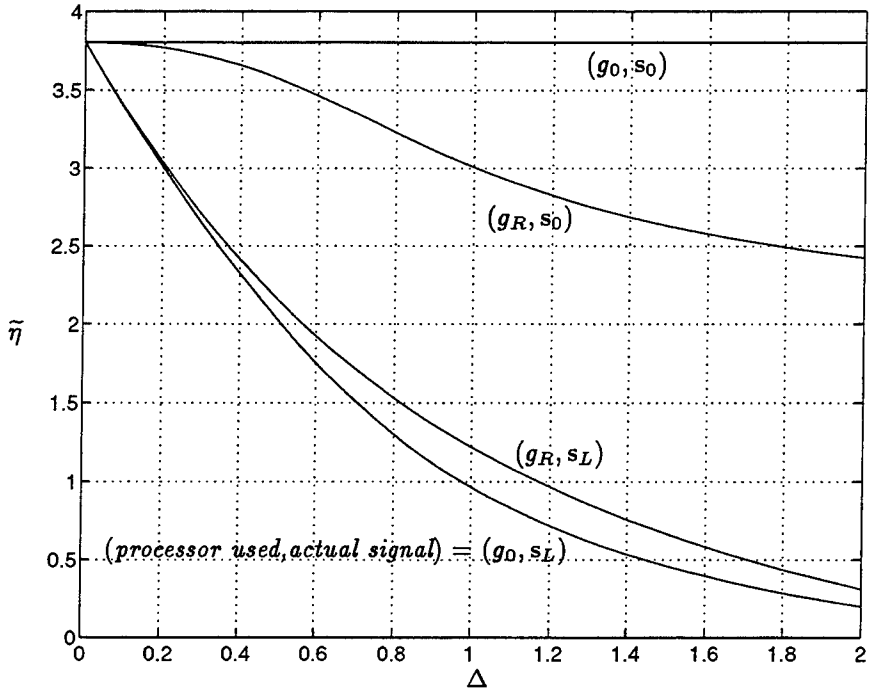

Figure 4.4: Asymptotic performance computations for the signal class $\mathcal{S}_2$.

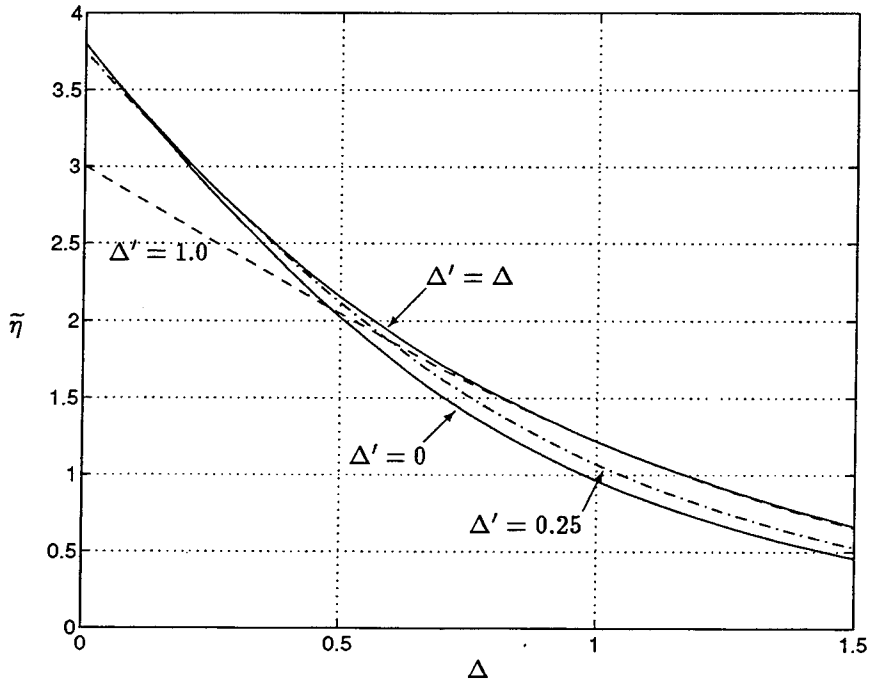

Figure 4.5: Design tradeoffs in selection of Δ' for the signal class $\mathcal{S}_2$.

and $x_2 \leq 1.5$. Let

$$A = \begin{pmatrix} -1 & 0 \\ 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} -2.5 \\ 3.5 \\ -0.5 \\ 1.5 \end{pmatrix}$$

Then the uncertainty region is equivalently described by the inequality $A\mathbf{x} \leq \mathbf{b}$, where $\mathbf{x} \geq \mathbf{0}$. In general, when $\mathbf{x} \in \Re^k$, the uncertainty region is described by the intersection of $2p$ hyperplanes. It is not difficult to see that $\mathcal{S}_1$ can also be described in this manner.

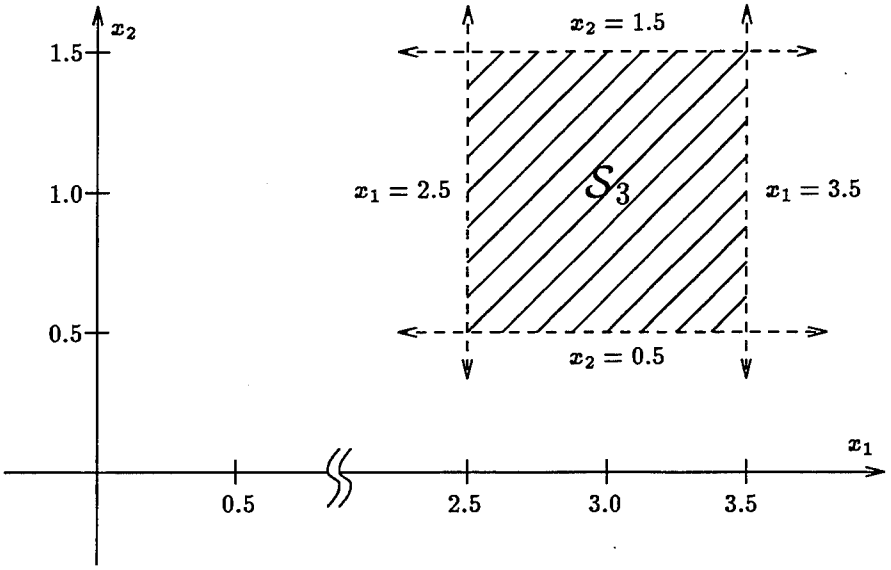

Figure 4.6: Signal uncertainty class $\mathcal{S}_3$.

As discussed earlier in this section, the robust processor is simply the log-likelihood ratio for the least favorable distributions, namely, those whose mean vectors are $\pm\mathbf{s}_L$. It was also shown that $\mathbf{s}_L$ is chosen to minimize the K-L divergence, $I(f_1, f_0) = 2\mathbf{s}^T\Sigma_0^{-1}\mathbf{s}$, where $\mathbf{s} \in \mathcal{S}$. Therefore, the least favorable signal can be obtained by

solving the following constrained minimization problem:

$$\text{minimize} \quad \mathbf{s}^T \Sigma_0^{-1} \mathbf{s}$$

$$\text{subject to} \quad A\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0}$$

This is the well-known *quadratic programming* problem, which has been studied thoroughly. In [2], it is shown that this problem can be solved using the *modified simplex method* by applying the *Karush-Kuhn-Tucker conditions*, which are necessary and sufficient for an optimal solution to exist. In [5], the solution is determined by applying a descent procedure to the *dual* problem. The details of these approaches can be found in the references listed.

4.3.2 Noise Covariance Uncertainty

In this section, we assume that the signal $\mathbf{s}_0$ is fixed and known, but that the noise covariance Σ lies in the uncertainty class $\mathcal{N}$. This situation might arise in a detection scheme designed to indicate the presence of a disorder, where the observables, $\mathbf{x}$, are snapshots obtained via multiple sensors. As discussed previously, if the sensor covariance, Σ, is known, the optimal processor, $g(\mathbf{x})$, is the log-likelihood ratio. On the other hand, when Σ is not known, a natural approach is to investigate alternative robust processors.

Types of Noise Uncertainty

In [10], two noise classes of the form

$$\mathcal{N} = \{\Sigma : \|\Sigma - \Sigma_0\| \leq \varepsilon, \Sigma > 0\}$$

are considered:

- $\mathcal{N}_1 = \{\Sigma : \|\Sigma - \Sigma_0\| \leq \varepsilon, \Sigma > 0\}$, *where* $\|\cdot\|$ *is the unit matrix norm; that is, any norm with the property that* $\|I\| = 1$

$$\begin{aligned} \mathbf{h}_L &= \Sigma_L^{-1}\mathbf{s}_0 \\ \Sigma_L &= \Sigma_0 + \varepsilon I \end{aligned}$$

- $\mathcal{N}_2 = \{\Sigma : \|\Sigma - \Sigma_0\| \leq \varepsilon, \Sigma > 0\}$, $\|\cdot\|$ *is the Euclidean norm; that is,* $\|A\|_2^2 = \sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\left[(A)_{ij}^2\right]$

$$\begin{aligned} \mathbf{h}_L &= (\Sigma_0 + \varepsilon I)^{-1}\mathbf{s}_0 \\ \Sigma_L &= \Sigma_0 + \sigma_n^2 \mathbf{h}_L \mathbf{h}_L^T \\ \varepsilon &= \sigma_n^2 \|\mathbf{h}_L\|^2 \end{aligned}$$

The above noise model can arise in a variety of applications, such as in radar or sonar problems where the signal to be detected is embedded in noise which is not completely characterized. For example, a typical underwater environment consists of uniform ambient background noise, superimposed with other sources such as impulsive or nonstationary noise components, and possibly some additional signals which are not of central interest. In such an environment, the sensor covariance, Σ, can be modelled as follows. Let

$$\Sigma = \Sigma_k + \Sigma_u$$

Here Σ_k is the "known" (or adequately estimated) component of the covariance, consisting of the contributions of the uniform background noise as well as measurement uncertainty (usually taken to be i.i.d.), while Σ_u is the "unknown" component, which accounts for all of the other sources of interference. The robust techniques discussed above can be directly applied here by defining the noise uncertainty class

$$\mathcal{N} = \{\Sigma : \|\Sigma - \Sigma_k\| \leq \varepsilon\}$$

where ε is selected sufficiently large to account for possible values of Σ_u.

Asymptotic Performance Under Noise Covariance Uncertainty

Denote the assumed and true noise covariance matrices as Σ_1 and Σ_2, respectively. The procedure for computing $\tilde{\eta}$, given below, is similar to that used in the case of signal uncertainty.

Suppose Page's test is implemented using $g_1(\mathbf{x}) = \mathbf{h}_1^T\mathbf{x}$, where $\mathbf{h}_1 \triangleq \Sigma_1^{-1}\mathbf{s}_0$. Define $\mathbf{v} \triangleq \omega_0\mathbf{h}_1$. The moment generating function equality (4.5) in this case is

$$\begin{aligned} 1 &= \int_{\Re^k} \exp\{\mathbf{v}^T\mathbf{x}\}(2\pi\Sigma_2)^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}+\mathbf{s}_0)^T\Sigma_2^{-1}(\mathbf{x}+\mathbf{s}_0)\right\}d\mathbf{x} \\ &= \exp\left\{-\mathbf{v}^T\mathbf{s}_0 + \frac{1}{2}\mathbf{v}^T\Sigma_2\mathbf{v}\right\} \end{aligned}$$

where the last expression is just the multivariate Gaussian moment generating function. Taking the log of both sides, substituting for $\mathbf{v}$, and solving for ω_0, we get

$$\omega_0 = \frac{2\mathbf{h}_1^T\mathbf{s}_0}{\mathbf{h}_1^T\Sigma_2\mathbf{h}_1}$$

Also noting that $\mathrm{E}\{g_1(x) \mid H_1\} = \mathbf{h}_1^T\mathbf{s}_0$, we have

$$\tilde{\eta}(g_1; \mathbf{s}_0, \Sigma_2) = \omega_0\mathrm{E}\{g_1(x) \mid H_1\} = 2\frac{|\langle \mathbf{h}_1, \mathbf{s}_0\rangle|^2}{\langle \mathbf{h}_1, \Sigma_2\mathbf{h}_1\rangle} \tag{4.12}$$

When the noise assumption is correct (i.e., $\Sigma_1 = \Sigma_2$), $\tilde{\eta} = \eta$ as discussed previously, and then

$$\tilde{\eta}(g_1; \mathbf{s}_0, \Sigma_1) = 2\mathbf{s}_0^T\Sigma_1^{-1}\mathbf{s}_0$$

The performance of the nominal and robust procedures can be evaluated by replacing Σ_1 with Σ_0 and Σ_L, respectively.

Example

Below is an example illustrating the procedure for determining the robust processor under covariance uncertainty. As in the example of the previous section, the nominal

parameters are

$$\mathbf{s}_0 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \quad \text{and} \quad \Sigma_0 = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$$

and now assume that the covariance uncertainty parameter is $\varepsilon = 2$.

- $\underline{\Sigma \in \mathcal{N}_1}$

  We have directly

$$\Sigma_L = \begin{pmatrix} 10+\varepsilon & 0 \\ 0 & 1+\varepsilon \end{pmatrix} = \begin{pmatrix} 12 & 0 \\ 0 & 3 \end{pmatrix}$$

  and

$$\mathbf{h}_L = \Sigma_L^{-1}\mathbf{s}_0 = \begin{pmatrix} 0.25 \\ 0.3\bar{3} \end{pmatrix}$$

  As with the signal distortion case, the asymptotic performance can be computed for the robust and nominal procedures. For the least favorable covariance, we have

$$\tilde{\eta}(g_0; \mathbf{s}_0, \Sigma_L) = 2\frac{\mid \mathbf{s}_0^T\Sigma_0^{-1}\mathbf{s}_0 \mid^2}{\mathbf{s}_0^T\Sigma_0^{-1}\Sigma_2\Sigma_0^{-1}\mathbf{s}_0} = 1.770$$

  and

$$\tilde{\eta}(g_R; \mathbf{s}_0, \Sigma_L) = 2\mathbf{s}_0^T\Sigma_L^{-1}\mathbf{s} = 2.167$$

  As expected, the robust quickest detector outperforms the nominal version when the covariance is least favorable.

- $\underline{\Sigma \in \mathcal{N}_2}$

  First, we have

$$\mathbf{h}_L = \begin{pmatrix} 12 & 0 \\ 0 & 3 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.25 \\ 0.3\bar{3} \end{pmatrix}$$

  Now

$$\|\mathbf{h}_L\|_2^2 = \langle \mathbf{h}_L, \mathbf{h}_L \rangle = 0.1736$$

and so

$$\sigma_n^2 = \frac{\varepsilon}{\|\mathbf{h}_L\|_2^2} = 11.52$$

Also

$$\mathbf{h}_L\mathbf{h}_L^T = \begin{pmatrix} 0.0625 & 0.083\bar{3} \\ 0.083\bar{3} & 0.1\bar{1} \end{pmatrix}$$

Therefore,

$$\Sigma_L = \Sigma_0 + \sigma_n^2\mathbf{h}_L\mathbf{h}_L^T = \begin{pmatrix} 10.72 & 0.96 \\ 0.96 & 2.28 \end{pmatrix}$$

The asymptotic performance for the nominal processor is

$$\widetilde{\eta}(g_0; \mathbf{s}_0, \Sigma_L) = 1.890$$

and for the robust version is

$$\widetilde{\eta}(g_R; \mathbf{s}_0, \Sigma_L) = 2.167$$

In Figure 4.7, the effect of mismatch in the assumed and true covariance contamination for the class $\mathcal{N}_2$ is examined; let ε' and ε, respectively, denote these quantities. The solid lines indicate the asymptotic performance when the nominal procedure ($\varepsilon' = 0$) and the optimal robust procedure ($\varepsilon' = \varepsilon$) are implemented. Observe that the test designed for a high uncertainty of $\varepsilon' = 5$ suffers a significant loss of performance if ε is small; in particular, for $\varepsilon < 1$, it would be better to use the nominal procedure. The small distortion performance can be improved by instead selecting $\varepsilon' = 1$, but this is at the expense of performance when the distortion level is high. Thus, analogous to the signal distortion example of the previous section, we conclude that a small amount of error in selecting ε' is tolerable.

4.3.3 Signal and Noise Uncertainty

In this section, we discuss the problem of designing the robust procedure for the case when uncertainty lies in both the signal and noise covariance; that is, $(\mathbf{s}, \Sigma) \in \mathcal{S} \times \mathcal{N}$.

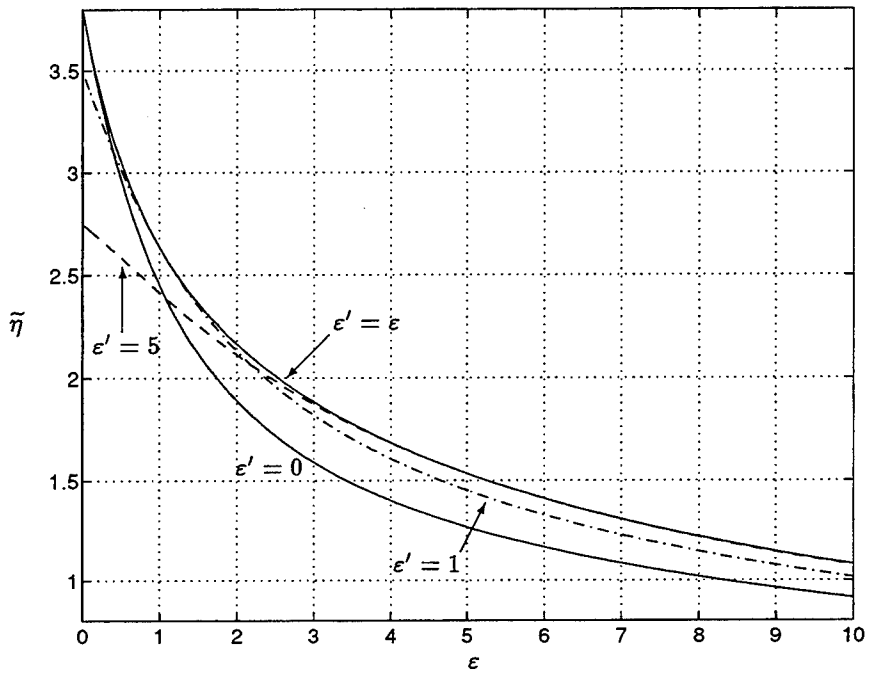

Figure 4.7: Design tradeoffs in selection of ε' for the signal class $\mathcal{N}_2$.

In general, the robust filter $\mathbf{h}_R$ and the least favorable pair $(\mathbf{s}_L, \Sigma_L)$ can be obtained by applying the three conditions of Lemma 1 in Section 4.2.

Since the classes $\mathcal{S}$ and $\mathcal{N}$ are independent, $\mathbf{s}_L$ and Σ_L can be determined separately whenever they can be written independently of $\mathbf{h}_R$. The robust processor can be obtained in two steps. First, determine the least favorable covariance Σ_L. Second, determine the least favorable signal $\mathbf{s}_L$ for the nominal parameter pair $(\mathbf{s}_0, \Sigma_L)$. For example, for the case when $(\mathbf{s}, \Sigma) \in \mathcal{S}_2 \times \mathcal{N}_2$, the robust processor $g_R(\mathbf{x}) = \mathbf{h}_R^T \mathbf{x}$ where

$$\mathbf{h}_L = \left(\Sigma_0 + (\varepsilon + \sigma_s^2) I\right)^{-1} \mathbf{s}_0$$

with $\Delta = \sigma_s^2 \parallel h_L \parallel$. For this case, it is interesting to notice that the robust processor accounts for both the signal and noise uncertainty by adding a white noise component (proportional to Δ and ϵ) to the nominal covariance Σ_0.

Suppose that the parameters are assumed to be $(\mathbf{s}_1, \Sigma_1)$, but in fact they are

$(\mathbf{s}_2, \Sigma_2)$. Thus, the test that is implemented uses the processor $g_1(\mathbf{x}) = \mathbf{h}_1^T\mathbf{x}$, where $\mathbf{h}_1 \triangleq \Sigma_1^{-1}\mathbf{s}_1$. Define $\mathbf{v} \triangleq \omega_0\mathbf{h}_1$. The left hand side of (4.5) is now

$$\int_{\Re^k} \exp\{\mathbf{v}^T\mathbf{x}\}(2\pi\Sigma_2)^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}+\mathbf{s}_2)^T\Sigma_2^{-1}(\mathbf{x}+\mathbf{s}_2)\right\}d\mathbf{x} = \exp\left\{-\mathbf{v}^T\mathbf{s}_2 + \frac{1}{2}\mathbf{v}^T\Sigma_2\mathbf{v}\right\}$$

Solving for the non-zero root, we get

$$\omega_0 = \frac{2\mathbf{h}_1^T\mathbf{s}_2}{\mathbf{h}_1^T\Sigma_2\mathbf{h}_1}$$

The other term is $\mathrm{E}\{g_1(x) \mid f\} = \mathbf{h}_1^T\mathbf{s}_2$; therefore

$$\widetilde{\eta}(g_1; (\mathbf{s}_2, \Sigma_2)) = 2\frac{|\langle \mathbf{h}_1, \mathbf{s}_2\rangle|^2}{\langle \mathbf{h}_1, \Sigma_2\mathbf{h}_1\rangle}$$

4.4 Extensions

4.4.1 Relationship to the "Minimax Tuning" Approach

In the preceding sections, the minimax criterion was directly applied to the approximate asymptotic performance measure $\widetilde{\eta}$, and the optimal processor was shown to be the minimax robust matched filter. In this section, we briefly discuss an alternative method for deriving the robust quickest detector. It turns out that the solutions for both problems are the same, although the approaches differ.

Suppose that we wish to implement Page's test using the linear processor $g(\mathbf{x}) = \mathbf{s}^T\Sigma^{-1}\mathbf{x}$, and notice that this is just the log-likelihood ratio for testing between $\mathcal{N}(-\mathbf{s}, \Sigma)$ and $\mathcal{N}(\mathbf{s}, \Sigma)$. [6] Thus, the test is designed assuming that $(\mathbf{s}, \Sigma)$ are the true parameters. However, suppose that the actual mean vector is $\boldsymbol{\theta}$. Recall that in Chapter 2, we defined the average sample number (ASN) of a CUSUM procedure with initial score z to be $\mathcal{N}_z(\boldsymbol{\theta})$. In [1], the following Wald approximations to the ASN are given.

$$\mathcal{N}_0(\boldsymbol{\theta}) \approx \widehat{\mathcal{N}}_0(\boldsymbol{\theta}) = \frac{e^{-\frac{2\mu}{\sigma^2}h} - 1 + \frac{2\mu}{\sigma^2}h}{\frac{2\mu^2}{\sigma^2}}, \quad \boldsymbol{\theta} \neq \mathbf{0}$$

[6] The constant multiplier has been omitted for convenience.

$$\mathcal{N}_0(\mathbf{0}) \approx \widehat{\mathcal{N}}_0(\mathbf{0}) = \frac{h^2}{\sigma^2}$$

where $\mu = \mathrm{E}[g(\mathbf{x}) \mid \boldsymbol{\theta}]$ and $\sigma^2 = \mathrm{Var}[g(\mathbf{x}) \mid \boldsymbol{\theta}]$. It is straightforward to show that $\mu = \mathbf{s}^T\Sigma^{-1}\boldsymbol{\theta}$ and $\sigma^2 = \mathbf{s}^T\Sigma^{-1}\mathbf{s}$.

Define the parameter

$$b = \frac{\mu}{\sigma} = \frac{\mathbf{s}^T\Sigma^{-1}\boldsymbol{\theta}}{(\mathbf{s}^T\Sigma^{-1}\mathbf{s})^{\frac{1}{2}}}$$

Now $\widehat{\mathcal{N}}_0(\boldsymbol{\theta})$ and $\widehat{\mathcal{N}}_0(\mathbf{0})$ can be directly related as

$$\widehat{\mathcal{N}}_0(\boldsymbol{\theta}) \triangleq \widehat{\mathcal{N}}_0(b) = \frac{e^{-2b\widehat{\mathcal{N}}_0^{\frac{1}{2}}(\mathbf{0})} + 2b\widehat{\mathcal{N}}_0^{\frac{1}{2}}(\mathbf{0}) - 1}{2b^2}$$

Under H_1, $b > 0$ (since $\mu > 0$), and $\widehat{\mathcal{N}}_0(b)$ is the worst expected delay in detection when the disorder occurs; similarly, under H_0, $b < 0$, and $\widehat{\mathcal{N}}_0(b)$ is the mean time between false alarms. Since the goal is to minimize the former and maximize the latter, one would like to choose $\mathbf{s}$ such that b^2 is maximized: this is because when $b > 0$, the b^2 term in the denominator dominates $\widehat{\mathcal{N}}_0(b)$, while when $b < 0$, the exponential term in the numerator dominates. Notice that

$$b^2 = \frac{\left|\mathbf{s}^T\Sigma^{-1}\boldsymbol{\theta}\right|^2}{\mathbf{s}^T\Sigma^{-1}\mathbf{s}} = \frac{|\langle \mathbf{h}, \boldsymbol{\theta}\rangle|^2}{\langle \mathbf{h}, \mathbf{s}\rangle}\bigg|_{\Sigma\mathbf{h}=\mathbf{s}}$$

which is the same as the value of $\tilde{\eta}$ (less a factor of two). The minimax tuning approach [1] is to maximize b^2 for the least favorable mean vector $\boldsymbol{\theta}$. However, since maximizing b^2 is equivalent to maximizing $\tilde{\eta}$, the results from minimax robust matched filtering can also be applied to the minimax tuning approach to obtain specific solutions for the robust processor.

The above approach can also be extended to include the case of covariance uncertainty. As in Section 4.3.2, let Σ_1 and Σ_2 denote the assumed and true covariances, and suppose that the true signal, $\mathbf{s}_0$, is known. Page's test is then implemented using the processor $g(\mathbf{x}) = \mathbf{s}_0^T\Sigma_1^{-1}\mathbf{x}$. This results in $\mu = \mathbf{s}_0^T\Sigma_1^{-1}\mathbf{s}_0$ and $\sigma^2 = \mathbf{s}_0^T\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}\mathbf{s}_0$.

Thus,

$$b^2 = \frac{\left|\mathbf{s}_0^T \Sigma_1^{-1} \mathbf{s}_0\right|^2}{\mathbf{s}_0^T \Sigma_1^{-1} \Sigma_2 \Sigma_1^{-1} \mathbf{s}_0} = \left.\frac{|\langle \mathbf{h}_1, \mathbf{s}_0 \rangle|^2}{\langle \mathbf{h}_1, \Sigma_2 \mathbf{h}_1 \rangle}\right|_{\mathbf{h}=\Sigma_1^{-1}\mathbf{s}_0}$$

Again, the equivalence between maximizing b^2 and $\tilde{\eta}$ in equation (4.12) is apparent.

4.4.2 Computation of $\tilde{\eta}$ For Non-Gaussian Noise

As mentioned in Section 4.1, in some cases it may be desirable to implement a quickest detection procedure designed using a maximum SNR criterion even if the noise is non-Gaussian. Here, the computation of $\tilde{\eta}$ which appears in the proof of Proposition 1 is generalized to include this case.

Suppose $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ are multivariate non-Gaussian densities, and let $M_0(\mathbf{v})$ denote the moment generating function of $f_0(\mathbf{x})$. Now $\tilde{\eta}$ is obtained via (4.4) and (4.5) as before; these expressions are repeated here for convenience:

$$\tilde{\eta} = \omega_0 \mathrm{E}\left[g(\mathbf{x}) \mid H_1\right] \tag{4.13}$$

$$1 = \int e^{\mathbf{v}^T \mathbf{x}} f_0(\mathbf{x}) d\mathbf{x}, \quad \mathbf{v} \triangleq \omega_0 \mathbf{h}_R \tag{4.14}$$

Now observe that this last expression is just the vector moment generating function as a function of $\mathbf{v}$. Therefore, ω_0 is determined by solving the equation

$$M_0(\mathbf{v}) = M_0(\omega_0 \mathbf{h}_R) = 1$$

either directly or numerically, and then (4.13) can be determined in a straightforward manner.

4.5 Conclusions

In this chapter, robust quickest detection procedures were investigated for the case of multivariate Gaussian observables with uncertain means and/or covariances. This statistical model arises in several areas, including radar, sonar, and other multisensor applications.

The most significant contribution establishes a formal connection between robust quickest detection and robust matched filtering. This allows one to apply previous results on the latter to the quickest detection problem when uncertainty in the first and second order statistics exists. Explicit solutions for the discrete-time robust matched filter, derived in [10], were used to separately obtain the robust quickest detectors for uncertainty in the mean vector (signal) and noise covariance.

When the observables are Gaussian, the robust quickest detector is optimal in the sense that the asymptotic performance measure $\widetilde{\eta}$ is maximized for the least favorable mean and covariance. It was also pointed out that, when the noise is non-Gaussian, the same techniques may be used to obtain a robust detector, where the goal is simply to maximize the SNR over the uncertainty class. In either case, expressions for $\widetilde{\eta}$ were derived which characterize the worst case performance of the robust procedure.

Simple examples were given to illustrate the design process. In some cases, such as when the noise is uncorrelated, the design of the processor that is robust to both signal and noise uncertainty can be carried out by separately determining the least favorable mean and covariance. In the more general case, the solution can be obtained by iteratively solving the set of equations in Lemma 1.

There are several interesting directions for future work. First, the robust quickest detection problem could be further generalized in a Hilbert space setting, as is done in [8] for the robust matched filtering problem. Second, it would be useful to obtain the continuous time robust quickest detector. Finally, it would be useful to determine

explicit solutions for the signal distortion problem when the noise is correlated. It was shown that this problem can be reformulated as a quadratic programming problem, for which two iterative approaches were mentioned. A direct solution to this problem would be useful in both robust quickest detection and robust matched filtering, and so this area is worthy of additional attention.

4.6 References

[1] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] F. S. Hillier and G. J. Lieberman, *Introduction to Operations Research*, Holden-Day, Inc., Oakland, CA, 1986.

[3] S. A. Kassam, *Signal Detection in Non-Gaussian Noise*, Springer-Verlag, New York, 1988.

[4] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey", *Proc. of the IEEE*, vol. 73, no. 3, March 1985.

[5] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, Inc., New York, 1969.

[6] J. H. Miller, *Signal Detection in Non-Gaussian Noise*, Ph.D. Thesis, Princeton University, 1972.

[7] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100-114, 1954.

[8] H. V. Poor, "Robust matched filters," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, Sept. 1983.

[9] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed., Springer-Verlag, New York, 1994.

[10] S. Verdu and H. V. Poor, "Minimax robust discrete-time matched filters," *IEEE Trans. Commun.*, vol. COM-31, no. 2, Feb. 1983.

Chapter 5

Quickest Detection in Decentralized Decision Systems

5.1 Introduction

In recent years, there has been an increasing interest in the area of *distributed*, or *decentralized*, detection. A distributed detection system contains two basic entities. The first is a collection of *local detectors*, each of which consists of a sensor followed by some type of decision rule. The second is a central processor, or *fusion center*, which processes the local decisions and produces a final decision. The use of such decentralized decision schemes is motivated by the reduction in channel bandwidth that can be achieved (and hence a reduction in system cost), and also by the need in some situations for the sensors to be separated by great distances. In addition, a decrease in the complexity of the decision procedure at the fusion center can be realized. However, by reducing the data locally instead of utilizing the complete data set at the central processor, some performance is also sacrificed.

Early work in distributed detection focused on one-step procedures; that is, where the decision is based on a single finite sample. In [16], the classical Bayesian approach to detection theory is extended to the case of distributed detection. The optimization

of the local and fusion procedures is studied in [16, 5, 13], and is extended in [7, 10] to the case where the local decisions are correlated. More recently, the distributed detection problem which incorporates sequential schemes at the local detectors [8] and fusion center [17, 9, 18] have been considered.

In this chapter, we consider the problem of detecting disorders using a distributed system. To date, there has been little work in this area. In Teneketzis [14] and Teneketzis and Varaiya [15], the decentralized quickest detection problem is formulated by defining a Bayes cost function which penalizes false alarms before the disorder and large delays in detection after the disorder. The disorder is modelled using a Markov chain, where the conditional probability of the jump occurring at time $i+1$ given that it did not occur at time i is some fixed value, which is presumably known or inferred from previous data. It is shown in [15] that, for the case where the cost function is not separable with respect to the local decisions, the local thresholds are the solutions to a set of coupled dynamic programming equations and are time-varying. A separable cost function is also considered; this results in fixed local thresholds, although with a longer delay in detection.

In this work, we examine several fusion rules for the case where the local detectors simply consist of an integrator followed by a comparator with a fixed threshold. As stated in Chapter 2, the disorder time is taken to be unknown. Several of the fusion procedures we consider assume knowledge of the signal strengths before and after the disorder; however, we also examine some procedures which are applicable when this information is not available.

In Section 2, the decentralized detection problem is stated precisely, and the notation used here (in addition to that of Chapter 2) is presented. In Section 3, the three distributed procedures under consideration are derived. The first of these is the ML optimal test, which is shown to admit a recursive form. The second is a version of Page's test, similar to the above test, but which requires less computation at each

iteration. The third test is a procedure which is suitable for the important case where the magnitude of the disorder is unknown. In Section 4, we show how the Markov approximation approach of Chapter 2 can be modified to compute the performance of the distributed system. In Section 5, a simple procedure for choosing the thresholds for the local decision rules is derived based upon an asymptotic performance measure. It is shown that not only can the thresholds be easily computed, but that the resulting performance is optimal for practical purposes; the latter point is the subject of Section 7. In Section 6, the performance of each of the tests is computed for strong and weak jump magnitude scenarios. Finally, in Section 8, the choice of blocklength of the local detectors is investigated. It is shown that, in general, the more samples used in the local decisions, the lower the performance and channel bandwidth cost; however, in the small signal case, it is actually advantageous to use a larger blocksize from a performance standpoint. [1]

5.2 Problem Statement

The decentralized detection system under consideration is shown in Figure 5.1. Here $\{x_\ell(i)\}_{i=1}^n$ is the sequence of samples received by sensor ℓ up to time n, where $\ell = 1, 2, \ldots, L$, and L is the total number of sensors. The sampling frequency is fixed at $f_s = 1/T_s$. The disorder is modelled as a step change in the mean of the observables; that is:

$$H_0: \quad x_\ell(i) = n_\ell(i), \qquad i = 1, 2, \ldots, m-1$$

$$H_1: \quad x_\ell(i) = n_\ell(i) + s_\ell, \quad i = m, m+1, \ldots$$

for each sensor $\ell = 1, 2, \ldots, L$, where the $n_\ell(i)$ are samples from a zero mean Gaussian distribution which is both spatially and temporally uncorrelated with $\mathrm{E}[n_\ell^2(i)] = \sigma_\ell^2$, and m is the unknown *disorder time*. Note that although we have chosen a Gaussian

[1] A preliminary version of this work has appeared in [6].

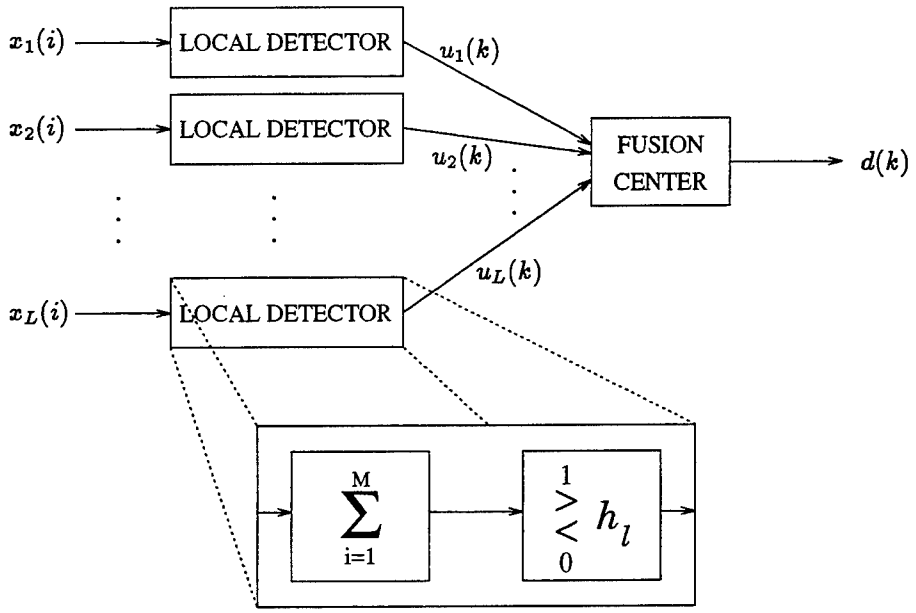

Figure 5.1: Structure of the distributed system.

noise model, the analyses which appear in subsequent sections are equally applicable when other distributions are used.

Each local detector consists of a summation block followed by a comparator as shown in Figure 5.1. A binary decision is made indicating whether or not the sum of M successive samples exceeds the fixed local threshold h_ℓ. The decisions are produced by the local detectors at a rate of one every M samples and are denoted by $\{u_\ell(k)\}$, where

$$u_\ell(k) = \mathcal{I}\{w_\ell(k) > h_\ell\} \tag{5.1}$$

$$w_\ell(k) = \sum_{j=M(k-1)+1}^{Mk} x_\ell(j) \tag{5.2}$$

for $\ell = 1, 2, \ldots, L$, where $\mathcal{I}\{\mathcal{A}\}$ is the indicator of the event $\mathcal{A}$. Specifically, sample $x_\ell(i)$ is involved in decision $u_\ell(k)$ if and only if $k = \lceil \frac{i}{M} \rceil$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Finally, global decisions $\{d(k)\}$ are produced at the fusion center based upon past and present local decisions.

The above local detection procedure is illustrated in Figure 5.2 for a blocksize of $M = 8$. Here k_0 denotes the block in which the disorder (at time sample m) occurs. The joint distributions of the samples in blocks $1, 2, \ldots, k_0 - 1$ are independent and identically distributed, as are those of blocks $k_0 + 1, k_0 + 2, \ldots$. Since the sample statistics are known before and after the disorder, the distributions of every block except k_0 are also known. In block k_0, the samples may have any one of M joint distributions due to the M possible disorder times within the block. [2]

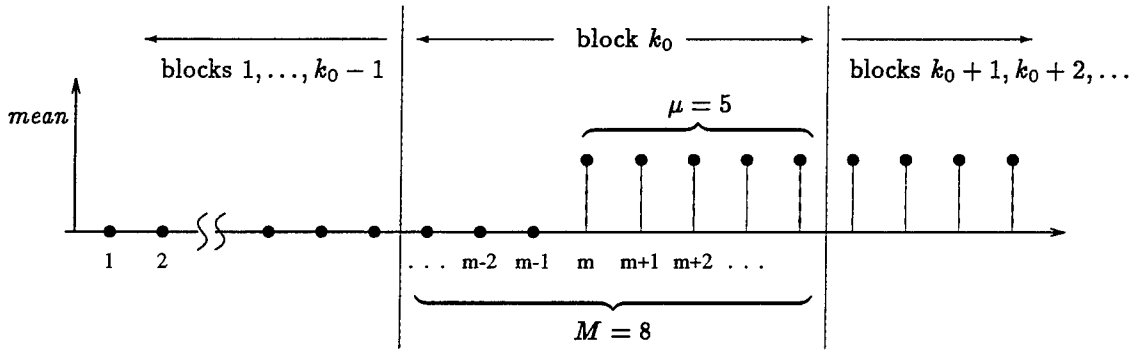

Figure 5.2: Operation of the local detectors.

As explained in Chapter 2, the goal of the overall procedure is to minimize the worst expected time to detect the disorder, D, subject to a lower bound on the mean time between false alarms, T. Along these lines, several options for the fusion rule will be considered, and their relative performance will be determined by computing plots of D versus $\log T$ for each. We will again be interested in the asymptotic performance measure

$$\eta = \lim_{T \to \infty} \frac{\log T}{D}$$

It will be shown in Section 5 that a lower bound $\tilde{\eta} \leq \eta$ can also be obtained for

[2] A slightly modified version of the above detection scheme can also be used in continuous time applications, such as in the case where multiple sensors are used to measure radar returns. The approach for the continuous-time problem is essentially the same, and is outlined in Appendix A.

the distributed detection problem and that this bound can be used to compute the thresholds of the local tests.

In analyzing the various distributed detection procedures, it will be convenient to utilize two versions of the usual stopping variable: [3]

- N is the stopping time expressed in *samples*
- $\widetilde{N}$ is the stopping time expressed in *blocks*

Since global decisions are produced only after each block of M snapshots is processed, it is natural to express T and D as the "expected numbers of blocks" before an alarm. However, since a disorder can occur at any one of M time instants within block k_0, we would ultimately like T and D to be in terms of samples (i.e., number of snapshots). This also allows us to compare procedures whose blocksizes differ.

Recall in Chapter 2, Section 4, that

$$T = \mathcal{N}_0(\theta_0) \text{ and } D = \mathcal{N}_0(\theta_1)$$

where $\mathcal{N}_z(\theta)$ was the ASN of Page's procedure with initial score z. Here we introduce a new definition, the *average block number* (ABN), which describes the stopping time of Page's test in terms of the number of blocks rather than number of samples. The two ABN's of interest are:

$\mathrm{E}_0\widetilde{N}$ $\triangleq$ expected number of blocks before stopping when all M samples in every block are generated under H_0

$\mathrm{E}_1\left[\widetilde{N} \mid \mu\right]$ $\triangleq$ expected number of blocks before stopping where:

i.) in the first block, $M - \mu$ samples are generated under H_0 and μ are generated under H_1

ii.) in all subsequent blocks, all of the M samples are generated under H_1

[3] Throughout this chapter, a tilde indicates units of blocks.

where the initial score is zero in both cases. In Appendix B, the following relationships between the ASN's and ABN's are derived:

$$\mathcal{N}_0(\theta_0) = M \cdot \mathrm{E}_0 \widetilde{N} \tag{5.3}$$

$$\mathcal{N}_0(\theta_1) = \mu + \left(\mathrm{E}_1 \left[\widetilde{N} \mid \mu\right] - 1\right) M \tag{5.4}$$

where again μ is the number of samples in block k_0 taken from H_1. Here we will let μ be fixed, but later we will take into account the fact that μ is actually random. To compute the performance of the various procedures, the ABN's will first be computed, and then converted to units of samples via (5.3)-(5.4).

5.3 Derivation of Fusion Rules

In this section, we introduce four procedures. First, the optimal ML procedure at the fusion center is derived, and it is shown that this test admits a recursive implementation. Second, a version of Page's test is considered; this procedure is more practical because it eliminates the need for performing an explicit maximization at every stage of the test. Third, a procedure that is suitable for the important case where $s_1, \ldots, s_L$ are unknown is presented; this procedure is sometimes referred to as Hinkley's test [1]. Finally, the ML optimal test for the case where all of the sensor data is available to the central processor is derived; the performance of this procedure is used as a standard to which the distributed procedures can be compared.

5.3.1 Known Signal Case

Let $\mathbf{u}(k) = \{u_\ell(k)\}_{\ell=1}^L \in \{0,1\}^L$ denote the local decisions for block k, and let $f(\mathbf{u} \mid \rho)$ denote the distribution of $\mathbf{u}$ given $\rho \in \{0, \ldots, M\}$, where the first $M - \rho$ samples of the block are from H_0 and the last ρ are from H_1. Thus, the decisions are distributed as $f(\mathbf{u} \mid \rho = 0)$ for blocks $1, \ldots, k_0 - 1$ and as $f(\mathbf{u} \mid \rho = M)$ for blocks $k_0 + 1, k_0 + 2, \ldots,$

while at block k_0 the distribution is $f(\mathbf{u}\,|\,\rho=\mu)$, for some $\mu \in \{1,\ldots,M\}$. Each local decision is computed according to (5.1)-(5.2). Since the observables are Gaussian, so are the sums $w_\ell(k)$. Specifically, if the joint distribution for block k is $f(\mathbf{u}\,|\,\rho)$, then $w_\ell(k) \sim \mathcal{N}(\rho s_\ell, M\sigma_\ell^2), \ell = 1,\ldots,L$, and $\Pr\{u_\ell(k)=1\} = \theta_\ell(\rho)$, where

$$\theta_\ell(\rho) = 1 - \Phi\left(\frac{h_\ell - \rho s_\ell}{\sqrt{M}\sigma_\ell}\right) \tag{5.5}$$

is the power of the local fixed sample test when the disorder occurs ρ samples before the end of the block, and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. [4]

The distributed quickest detection problem may be alternatively stated in terms of blocks rather than samples as follows. Define the hypotheses K_0, K'_μ, and K_1 as:

$$\begin{aligned}
K_0: &\quad \Pr\{u_\ell(k)=1\} = \alpha_\ell, && k = 1,\ldots,k_0-1 \\
K'_\mu: &\quad \Pr\{u_\ell(k)=1\} = \theta_\ell(\mu), && k = k_0,\ \ \mu \in \{1,2,\ldots,M\} \\
K_1: &\quad \Pr\{u_\ell(k)=1\} = \beta_\ell, && k = k_0+1, k_0+2,\ldots
\end{aligned} \tag{5.6}$$

where for convenience we have defined $\alpha_\ell \triangleq \theta_\ell(0)$ and $\beta_\ell \triangleq \theta_\ell(M)$. Thus, when a disorder occurs the progression $K_0 \rightarrow K'_\mu \rightarrow K_1$ results. [5]

Let $\mathcal{L}_n(\mu, k_0)$ denote the likelihood ratio of the local decisions up to and including block n assuming that the disorder occurs in block k_0, where $1 \le k_0 \le n$. Specifically:

$$\begin{aligned}
\mathcal{L}_n(\mu,k_0) &= \frac{\prod_{k=1}^{k_0-1} f(\mathbf{u}(k)\,|\,\rho=0)\cdot f(\mathbf{u}(k_0)\,|\,\rho=\mu)\cdot \prod_{k=k_0+1}^{n} f(\mathbf{u}(k)\,|\,\rho=M)}{\prod_{k=1}^{n} f(\mathbf{u}(k)\,|\,\rho=0)} \\
&= \frac{f(\mathbf{u}(k_0)\,|\,\rho=\mu)}{f(\mathbf{u}(k_0)\,|\,\rho=0)} \cdot \prod_{k=k_0+1}^{n} \frac{f(\mathbf{u}(k)\,|\,\rho=M)}{f(\mathbf{u}(k)\,|\,\rho=0)}, \quad 1 \le k_0 \le n
\end{aligned} \tag{5.7}$$

The ML procedure is to maximize (5.7) over the quantities μ and k_0; a disorder is declared if this quantity exceeds a fixed threshold. This is also called the generalized

[4] If the observations are not Gaussian, then $\theta_\ell(\rho)$ can be redefined accordingly.

[5] Observe that hypotheses K_0, K_1, and K'_μ are dependent on the hypotheses H_0 and H_1. In particular, K_0 holds if and only if every sample within the block is from H_0, and similarly for K_1 and H_1, while K'_μ holding implies that samples potentially come from both H_0 and H_1. Notice also that $K'_\mu = K_1$ when $\mu = M$.

likelihood ratio test (GLRT) [2]. In general, this test does not admit a recursive solution which would make the procedure more useful in real-time applications. However, a recursive implementation does exist for the present problem, as shown below.

Since the $u_\ell(k)$ are simply Bernoulli random variables, we have

$$f(\mathbf{u}\,|\,\rho) \;=\; \prod_{\ell=1}^{L}(\theta_\ell(\rho))^{u_\ell}(1-\theta_\ell(\rho))^{1-u_\ell} \tag{5.8}$$

Substituting (5.8) into (5.7) and taking the log of both sides, we have

$$\begin{aligned}
\ell_n(\mu,k_0) &\triangleq \log\mathcal{L}_n(\mu,k_0)\\
&= \log\prod_{\ell=1}^{L}\left(\frac{\theta_\ell(\mu)}{\alpha_\ell}\right)^{u_\ell(k_0)}\left(\frac{1-\theta_\ell(\mu)}{1-\alpha_\ell}\right)^{1-u_\ell(k_0)}\\
&\quad+\sum_{k=k_0+1}^{n}\log\prod_{\ell=1}^{L}\left(\frac{\beta_\ell}{\alpha_\ell}\right)^{u_\ell(k)}\left(\frac{1-\beta_\ell}{1-\alpha_\ell}\right)^{1-u_\ell(k)}\\
&= \sum_{\ell=1}^{L}\left\{u_\ell(k_0)\log\left(\frac{\theta_\ell(\mu)}{\alpha_\ell}\right)+(1-u_\ell(k_0))\log\left(\frac{1-\theta_\ell(\mu)}{1-\alpha_\ell}\right)\right\}\\
&\quad+(n-k_0)d+\sum_{k=k_0+1}^{n}\sum_{\ell=1}^{L}c_\ell u_\ell(k)
\end{aligned} \tag{5.9}$$

where $c_\ell \triangleq \log\left[\frac{\beta_\ell(1-\alpha_\ell)}{\alpha_\ell(1-\beta_\ell)}\right]$ and $d \triangleq \sum_{\ell=1}^{L}\log\left(\frac{1-\beta_\ell}{1-\alpha_\ell}\right)$. Define the test statistic

$$S_n = \max_{1\le k_0\le n}\max_{1\le\mu\le M}\ell_n(\mu,k_0)$$

The GLRT is then to declare a disorder at block n in case $S_n > h$, where the threshold h of the test is chosen to satisfy a false alarm condition. Since μ appears only in the first term of (5.9), we define

$$\phi(\mathbf{u}(k)) = \max_{1\le\mu\le M}\sum_{\ell=1}^{L}\left\{u_\ell(k)\log\left(\frac{\theta_\ell(\mu)}{\alpha_\ell}\right)+(1-u_\ell(k))\log\left(\frac{1-\theta_\ell(\mu)}{1-\alpha_\ell}\right)\right\} \tag{5.10}$$

for each block k, so that

$$S_n = \max_{1\le k_0\le n}\left\{\phi(\mathbf{u}(k_0))+(n-k_0)d+\sum_{k=k_0+1}^{n}\sum_{\ell=1}^{L}c_\ell u_\ell(k)\right\} \tag{5.11}$$

In order to obtain the recursive version of this test, assume that S_n has been computed prior to block $n+1$. Now:

$$\begin{aligned} S_{n+1} &= \max_{1\le k_0\le n+1}\left\{\phi(\mathbf{u}(k_0)) + (n+1-k_0)d + \sum_{k=k_0+1}^{n+1}\sum_{\ell=1}^{L} c_\ell u_\ell(k)\right\} \\ &= \max\left\{\phi(\mathbf{u}(n+1)),\ \max_{1\le k_0\le n}\left[\phi(\mathbf{u}(k_0)) + (n+1-k_0)d + \sum_{k=k_0+1}^{n+1}\sum_{\ell=1}^{L} c_\ell u_\ell(k)\right]\right\} \\ &= \max\left\{\phi(\mathbf{u}(n+1)),\ \max_{1\le k_0\le n}\left[\phi(\mathbf{u}(k_0)) + (n-k_0)d + \sum_{k=k_0+1}^{n}\sum_{\ell=1}^{L} c_\ell u_\ell(k)\right]\right. \\ &\qquad \left. + d + \sum_{\ell=1}^{L} c_\ell u_\ell(n+1)\right\} \end{aligned} \tag{5.12}$$

where in the last line the maximization has been computed separately over the sets $\{1\le k_0\le n\}$ and $\{k_0 = n+1\}$, and with the convention that $\sum_{i=j}^{k} \equiv 0$ when $j > k$. Finally, (5.11) and (5.12) together yield the recursive form of the test:

$$\mathcal{P}_1: \quad \begin{cases} S_n &= \max\{S_{n-1} + g_1(\mathbf{u}(n)), \phi(\mathbf{u}(n))\}, \quad S_0 = 0 \\ \widetilde{N}_1 &= \inf\{n \mid S_n > h\} \end{cases}$$

where

$$g_1(\mathbf{u}(n)) \triangleq \sum_{\ell=1}^{L} c_\ell u_\ell(n) + d$$

Here, $\widetilde{N}_1$ is the stopping time of the test and g_1 is the log-likelihood ratio for testing K_0 versus K_1. Note that the stopping time is expressed in blocks; the conversion to samples is done using (5.3)-(5.4). A block diagram of this procedure is shown in Figure 5.3.

One drawback of the optimal test is the need to compute $\phi(\mathbf{u}(k))$ for each block. Since this requires a maximization over a potentially large number of points, we are motivated to consider the following procedure:

$$\mathcal{P}_2: \quad \begin{cases} S_n &= \max\{S_{n-1} + g_2(\mathbf{u}(n)), 0\}, \quad S_0 = 0 \\ \widetilde{N}_2 &= \inf\{n \mid S_n > h\} \end{cases}$$

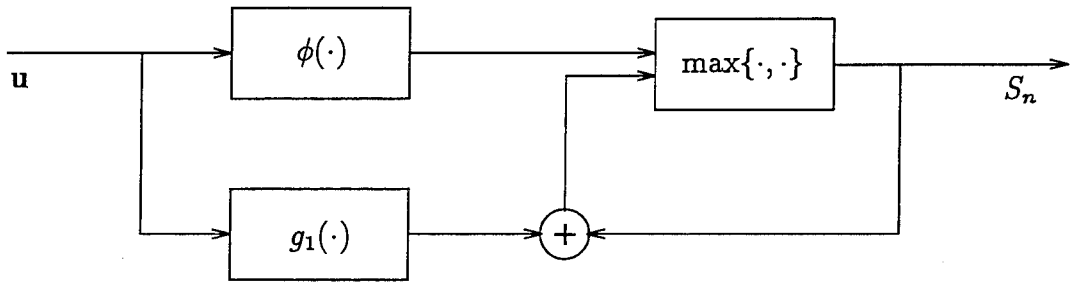

Figure 5.3: Structure of the ML optimal procedure.

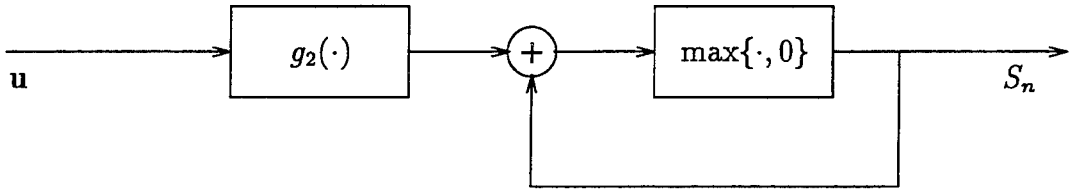

Figure 5.4: Structure of the suboptimal procedure.

where $g_2 \equiv g_1$ and $\phi(\cdot)$ is replaced by zero. The structure of this test is shown in Figure 5.4. This is just the familiar Page procedure of Chapter 2. Unlike procedure $\mathcal{P}_1$, $\mathcal{P}_2$ does not explicitly incorporate the information for the change block into the global decision. Note that $\mathcal{P}_1$ can be essentially viewed as a Page procedure with a lower boundary which is dependent on the data (the lower boundary is zero for $\mathcal{P}_2$). Another interesting point is that if it were known *a priori* that the disorder occurred at the beginning of a block, i.e. $\mu = M$, $\mathcal{P}_2$ would be the optimal test not only in the ML sense, but also in the sense of minimizing D for any fixed T, the criterion of Lorden which was discussed in Chapter 2.

5.3.2 Unknown Signal Case

In the above procedures, the jump magnitudes of the signals at the disorder time are taken to be known. However, in some situations, such as when the location of the phenomenon causing the disorder is not known, the resulting signal strengths will

also not be known. This motivates us to consider an additional version of Page's test suitable for problems where the jump magnitude is not known. The approach is similar to the case where the jumps $\{s_\ell\}_{\ell=1}^L$ are known, except that now we assume that $s_\ell \geq \delta_\ell$ for $\ell = 1, \ldots, L$, where $\{\delta_\ell\}_{\ell=1}^L$ is a set of minimum jump magnitudes. The derivations in Section 5.3.1 are then carried out assuming $s_\ell = \delta_\ell$. Thus, this test is designed for the disorder of minimum magnitude, although the monotonicity of the likelihood ratio means that the procedure will also react to larger jumps.

First, we assume without loss of generality that $\delta_\ell \equiv \delta$. This implies $\theta_\ell(\mu) = \theta(\mu)$, $\alpha_\ell = \alpha$, and $\beta_\ell = \beta$. Not only does this simplify the calculations, but it may also be a realistic assumption when little is known about the origin of the disturbance. Define $v(k) = \sum_{\ell=1}^L u_\ell(k)$. The log-likelihood ratio in (5.9) now reduces to

$$\begin{aligned} \ell_n(\mu, k_0) &= v(k_0) \log\left(\frac{\theta(\mu)}{\alpha}\right) + (L - v(k_0)) \log\left(\frac{1-\theta(\mu)}{1-\alpha}\right) \\ &\quad + (n - k_0)d + c \sum_{k=k_0+1}^{n} v(k) \end{aligned} \tag{5.13}$$

where $c \stackrel{\Delta}{=} \log\left[\frac{\beta(1-\alpha)}{\alpha(1-\beta)}\right]$ and $d \stackrel{\Delta}{=} L \log\left(\frac{1-\beta}{1-\alpha}\right)$. Following the same procedure as in (5.12) and again neglecting $\phi(\cdot)$, we eventually arrive at a sequential procedure similar to $\mathcal{P}_2$ for detecting changes of unknown signal strength:

$$\mathcal{P}_3 : \quad \begin{cases} S_n &= \max\{S_{n-1} + g_3(v(n)), 0\}, \quad S_0 = 0 \\ \widetilde{N}_3 &= \inf\{n \mid S_n \geq h\} \end{cases}$$

where $g_3(v(n)) = cv(n) + d$ is the log-likelihood ratio for testing K_1 versus K_0, where now K_1 denotes the minimum jump hypothesis (i.e., $s_\ell = \delta$, $\forall \ell$). We refer to this as Hinkley's test. The structure of $\mathcal{P}_3$ is similar to that of $\mathcal{P}_2$ (Figure 5.4), except that now the vector of local decisions $\mathbf{u}$ is replaced by the sum v. Notice that for this test, the fact that the minimum jump is δ for each sensor means that all measurements are equally weighted. Therefore, we are able to simply consider the sum of the sensor measurements for each snapshot.

5.3.3 Non-Distributed Case

In this case, all of the data is available at the fusion center (i.e., no local decisions are made). The ML optimal test, derived in Appendix C, is:

$$\mathcal{P}_0: \quad \begin{cases} S_n &= \max\{S_{n-1} + z_n, 0\}, \quad S_0 = 0 \\ N_0 &= \inf\{n \mid S_n > h\} \end{cases}$$

where

$$z_j \triangleq \sum_{\ell=1}^{L} \sigma_\ell^{-2} \left\{ x_\ell(j) s_\ell - \frac{1}{2} s_\ell^2 \right\}$$

This is the optimal procedure in terms of the criterion of Lorden (cf. Chapter 2). To see this, note that this procedure tests for a change in the mean of the univariate random variable z_j from $E(z_j \mid H_0)$ to $E(z_j \mid H_1)$, where

$$E(z_j \mid H_1) = -E(z_j \mid H_0) = \frac{1}{2} \sum_{\ell=1}^{L} \frac{s_\ell^2}{\sigma_\ell^2}$$

It will be shown that, while this test requires the largest channel bandwidth, it also yields the best performance because the information is not reduced locally. It is therefore included as a benchmark to which the other procedures are compared.

5.4 Performance Computation

For each of procedures $\mathcal{P}_0$ - $\mathcal{P}_3$, performance curves are generated by computing the pair (T, D) over a range of uniformly spaced values of h. For the non-distributed procedure $(\mathcal{P}_0)$, the Markov approximation method described in Chapter 2, Section 4 is used. A modified version of this method can also be used for the distributed procedures, as described below.

Recall from Chapter 2 that $r_i(n)$ was defined as the probability of reaching stage n under hypothesis H_i, and that the ASN could be expressed in terms of $\{r_i(n)\}_{i=1}^{\infty}$.

For the distributed procedures it is useful to redefine $r_i(n)$ in terms of blocks; that is, a "stage" in this case is just a block. Specifically, $\tilde{r}_0(n)$ is the probability of reaching block n when the disorder never occurs, and $\tilde{r}_1(n;\mu)$ is the probability when the disorder occurs at sample time μ in the *first* block.

The Markov approximation technique can be used as in Chapter 2 with one modification: the statistics of block k_0 (the change block) differ from those subsequent blocks. Therefore, it is necessary to compute separate probability transition matrices for each of the three hypotheses K_0, K'_μ, and K_1; denote these as $\mathbf{Q}_0, \mathbf{Q}_c$, and $\mathbf{Q}_1$, respectively, along with submatrices $\mathbf{R}_0$, $\mathbf{R}_c$, and $\mathbf{R}_1$. [6] At each stage, there are 2^L possible input vectors $\mathbf{u}$. Thus, the $\mathbf{Q}$'s can be determined in the following way. Define the states $\alpha_0, \ldots, \alpha_p$ and α^* as in Chapter 2, and let $\text{bin}(j)$ denote the binary version of integer j: for example, for $L = 4$ sensors, $\text{bin}(13) = [1\ 1\ 0\ 1]^T$. Let $\varphi(\cdot,\cdot)$ be a generic mapping for incrementing the test statistic; namely, if $\zeta_0, \zeta_1 \in \{\alpha_0, \ldots, \alpha_p, \alpha^*\}$, then $\zeta_1 = \varphi(\zeta_0, \mathbf{u})$ indicates that input $\mathbf{u}$ produces the state transition $\zeta_0 \mapsto \zeta_1$. Now $\mathbf{Q}$ can be determined using the following procedure:

1. Set $Q_{i,j} = 0,\ \forall i, j$

2. For $i = 0, 1, \ldots, p$ and $j = 0, 1, \ldots, 2^L - 1$:

   (a) $\mathbf{u} = \text{bin}(j)$, $p_u = \Pr\{\mathbf{u}\}$

   (b) $k = \ell$, where ℓ is such that $\alpha_\ell = \varphi(\alpha_i, \mathbf{u})$, with the convention that $k = p + 1$ if $\alpha^* = \varphi(\alpha_i, \mathbf{u})$

   (c) Set $Q_{i,k} \leftarrow Q_{i,k} + p_u$

3. Set $Q_{p+1,p+1} = 1$

[6] The subscript "c" stands for "change".

Step 3 reflects the fact that state $p+1$ is a terminal state.

The ABN's are determined in the same manner as are the ASN's in Chapter 2. Here, the expected stopping times are

$$\mathrm{E}_0\widetilde{N} = \sum_{n=1}^{\infty} n(\tilde{r}_0(n) - \tilde{r}_0(n+1)) = \sum_{n=1}^{\infty} \tilde{r}_0(n) \tag{5.14}$$

and similarly

$$\mathrm{E}_1\left[\widetilde{N} \mid \mu\right] = \sum_{n=1}^{\infty} \tilde{r}_1(n;\mu) \tag{5.15}$$

For the false alarm case, we have

$$\tilde{r}_0(n;\mu) = \pi_0' \mathbf{R}_0^{n-1} \underline{1},\ n = 1, 2, \ldots$$

which, when substituted into (5.14), results in

$$\mathrm{E}_0\widetilde{N} = \pi_0' \left(\mathbf{I} - \mathbf{R}_0\right)^{-1} \underline{1} \tag{5.16}$$

Notice that (5.16) is identical to (2.14) in Chapter 2. [7]

The computation of $\mathrm{E}_1\left[\widetilde{N} \mid \mu\right]$ is done in a similar manner, except that we must now also take the contribution of the change block into account. The transition matrix for this case is $\mathbf{Q}_c$ for the first stage and $\mathbf{Q}_1$ for subsequent stages. Thus

$$\begin{aligned}
\tilde{r}_1(1;\mu) &= 1 \\
\tilde{r}_1(2;\mu) &= \pi_0' \mathbf{R}_c \underline{1} \\
\tilde{r}_1(3;\mu) &= \pi_0' \mathbf{R}_c \mathbf{R}_1 \underline{1} \\
&\vdots \\
\tilde{r}_1(n;\mu) &= \pi_0' \mathbf{R}_c \mathbf{R}_1^{n-2} \underline{1}
\end{aligned} \tag{5.17}$$

Therefore, (5.15) together with (5.17) yields

$$\mathrm{E}_1\left[\widetilde{N} \mid \mu\right] = 1 + \sum_{n=2}^{\infty} \pi_0' \mathbf{R}_c \mathbf{R}_1^{n-2} \underline{1}$$

[7] Recall that π_0' is a truncated version of the state probability vector.

$$
\begin{aligned}
&= 1 + \pi_0' \mathbf{R}_c \left(\sum_{n=0}^{\infty} \mathbf{R}_1^n \right) \underline{1} \\
&= 1 + \pi_0' \mathbf{R}_c (\mathbf{I} - \mathbf{R}_1)^{-1} \underline{1} \qquad (5.18)
\end{aligned}
$$

Finally, the quantities in (5.16) and (5.18) are converted into units of samples via (5.3)-(5.4), which results in:

$$\mathcal{N}_0(\theta_0) = M \pi_0' (\mathbf{I} - \mathbf{R}_0)^{-1} \underline{1} \qquad (5.19)$$

$$\mathcal{N}_0(\theta_1) = \mu + M \pi_0' \mathbf{R}_c (\mathbf{I} - \mathbf{R}_1)^{-1} \underline{1} \qquad (5.20)$$

A different $\mathbf{R}_c$ is computed for each $\mu = 1, 2, \ldots, M$. However, observe that the product $\pi_0' \mathbf{R}_c$ simply picks off the first row of $\mathbf{R}_c$; therefore, for each μ, it is necessary only to compute the transition probabilities out of the initial state (α_0).

5.5 Choosing the Local Thresholds

The asymptotic performance measure, η, was defined in Chapter 2. It was seen that $\frac{1}{\eta}$ is the slope of the plot of D versus $\log T$ as $T \to \infty$, and therefore minimizing D corresponds (asymptotically) to maximizing η. It was also shown that the lower bound $\tilde{\eta} \leq \eta$ is useful because, for large T, it enables us to upper bound the worst expected delay as

$$D \leq \frac{\log T}{\tilde{\eta}}$$

and also it is not difficult to compute. The original derivation of $\tilde{\eta}$ appears in [4].

η and $\tilde{\eta}$ can be defined in the same manner for the distributed detection case with the caveat that the lower bound $\tilde{\eta}$ differs slightly for this problem. It turns out that for the ML optimal procedure, the lower bound is

$$\tilde{\eta} = \frac{1}{M} \omega_0 \mathrm{E}_1[g_1(\mathbf{u})] \qquad (5.21)$$

where ω_0 satisfies the moment generating function equality $\mathrm{E}_0[\exp\{\omega_0 g_1(\mathbf{u})\}] = 1$. The derivation of this bound is somewhat messy, and can be found in Appendix D.

It is also shown in this appendix that $\tilde{\eta}$ is the same for both $\mathcal{P}_1$ and $\mathcal{P}_2$. Thus, the optimal and suboptimal procedures are asymptotically equivalent. This means that for large values of T, there will be little difference in performance between the two procedures. The performance calculations in the next section corroborate this. Below, it is shown that $\tilde{\eta}$ can be useful in determining the local thresholds so as to optimize the asymptotic performance of the overall procedure.

Let $\mathbf{h}_{loc} = [h_1, h_2, \ldots, h_L]$ denote the vector of fixed local thresholds. Ideally, one would like to optimize the distributed system over all possible $\mathbf{h}_{loc} \in \Re_+^L$. In Chapter 2, we stated the objective of quickest detection: to minimize D subject to a lower bound on T. Using (5.19) and (5.20), this optimization problem can be written explicitly as:

$$\begin{aligned} &\textit{minimize} \quad \mu + M\pi_0' \mathbf{R}_c (\mathbf{I} - \mathbf{R}_1)^{-1} \underline{1} \;\; \textit{over } \mathbf{h}_{loc} \in \Re_+^L \\ &\textit{subject to} \quad M\pi_0' (\mathbf{I} - \mathbf{R}_0)^{-1} \underline{1} \geq c \end{aligned}$$

where c is a positive constant. This approach is difficult to implement for several reasons. The first reason is the existence of the change block. Here, the probability $\theta_\ell(\mu)$ of locally detecting the change at sensor ℓ is dependent on μ. Since μ is unknown, it is not possible to select an optimum threshold *a priori*. [8] Another difficulty is that the optimal $\mathbf{h}_{loc}$ is dependent on the desired mean time between false alarms, or similarly, on the desired global threshold h. One can see this by noting that a procedure with a higher h will produce an alarm later than the same one with a lower h; thus, the contribution of the change block to the final decision is more significant for lower h, and so in this case one would like to choose the local thresholds to take fuller advantage of the information extracted from this block. It is also not clear whether a unique solution to this nonlinear optimization problem exists. Since T is a function of both $\mathbf{h}_{loc}$ and h, fixing T does not lead to a fixed h, so h is also dependent on $\mathbf{h}_{loc}$. One

[8] One might consider assigning a uniform distribution to the arrival time within the change block and averaging over all possible μ.

might consider an adaptive search to maximize the performance over all $\mathbf{h}_{loc} \in \Re_+^L$. Unfortunately, the relatively long time required to compute the performance even for fixed thresholds would be prohibitive in an iterative scheme.

A simpler method for selecting the local thresholds based on the asymptotic lower bound $\tilde{\eta}$ in (5.21) is proposed here. The goal is to maximize $\tilde{\eta}$ over $\mathbf{h}_{loc} \in \Re_+^L$. When g is the log-likelihood ratio for choosing between K_0 and K_1, direct evaluation of the moment generating function identity [4] yields $\omega_0 = 1$. Since M is a constant, the goal is then

$$\max_{\mathbf{h}_{loc}} \mathrm{E}_1[g(\mathbf{u})] = \max_{\mathbf{h}_{loc}} \mathrm{E}_1 \left[\sum_{\ell=1}^{L} \left\{ u_\ell \log\left(\frac{\beta_\ell}{\alpha_\ell}\right) + (1-u_\ell)\log\left(\frac{1-\beta_\ell}{1-\alpha_\ell}\right) \right\} \right] \tag{5.22}$$

Since the local decisions are independent, the expectation can be applied termwise. Thus,

$$\begin{aligned} \max_{\mathbf{h}_{loc}} \mathrm{E}_1[g(\mathbf{u})] &= \max_{\mathbf{h}_{loc}} \sum_{\ell=1}^{L} \mathrm{E}_1 \left\{ u_\ell \log\left(\frac{\beta_\ell}{\alpha_\ell}\right) + (1-u_\ell)\log\left(\frac{1-\beta_\ell}{1-\alpha_\ell}\right) \right\} \\ &= \max_{\mathbf{h}_{loc}} \sum_{\ell=1}^{L} D_b(\beta_\ell, \alpha_\ell) \\ &= \sum_{\ell=1}^{L} \max_{h_\ell} D_b(\beta_\ell, \alpha_\ell) \end{aligned}$$

where we note that $\mathrm{E}_1[u_\ell] = \beta_\ell$ and $D_b(a,b) \triangleq a\log(\frac{a}{b}) + (1-a)\log(\frac{1-a}{1-b})$ is just the *binary discrimination* function from information theory [3]. Thus, the problem of globally maximizing the asymptotic lower bound amounts to selecting the local thresholds to maximize the binary discrimination for each sensor.

The local threshold h_ℓ is related to α_ℓ and β_ℓ for $\ell = 1, 2, \ldots, L$ via the equations:

$$\begin{aligned} \alpha_\ell &= \int_{h_\ell}^{\infty} \frac{1}{\sqrt{2\pi M}\sigma} \exp\left\{ -\frac{1}{2M\sigma^2}\tau^2 \right\} d\tau = 1 - \Phi\left(\frac{h_\ell}{\sqrt{M}\sigma}\right) \\ \beta_\ell &= \int_{h_\ell}^{\infty} \frac{1}{\sqrt{2\pi M}\sigma} \exp\left\{ -\frac{1}{2M\sigma^2}(\tau - \mu s_\ell)^2 \right\} d\tau = 1 - \Phi\left(\frac{h_\ell - \mu s_\ell}{\sqrt{M}\sigma}\right) \end{aligned}$$

The function $D_b(\beta_\ell(h_\ell), \alpha_\ell(h_\ell))$ has a unique maximum over $h_\ell \in [0, \infty]$, so the optimal threshold is easy to compute. The explicit solution satisfies a transcendental

equation; therefore, we instead use a binary search procedure allowing us to get arbitrarily close to the optimal threshold. The merits of using the above scheme for choosing the local thresholds will be evaluated in Section 7 via a sensitivity analysis.

As a sidenote, it is not difficult to show that regardless of the choice of $\mathbf{h}_{loc}$, the optimal non-distributed procedure is always asymptotically better than its distributed counterpart. Let η_{dis} and η_{non} denote the asymptotic performances for the distributed and non-distributed procedures, respectively. Because the log-likelihood ratio is used for each case, the lower bounds are tight. It is shown in Appendix C that

$$\eta_{non} = \sum_{\ell=1}^{L} \frac{s_\ell^2}{2\sigma_\ell^2}$$

Also, from the above discussion, when the optimum local thresholds are used:

$$\eta_{dis} = \tilde{\eta}_{dis} = \frac{1}{M} \sum_{\ell=1}^{L} \max_{h_\ell} D_b(\beta_\ell, \alpha_\ell) \tag{5.23}$$

The equality in (5.23) is strict since the log-likelihood ratio is used at the fusion center [4]. Now Corollary 4.4.2 in [3] provides the following bound:

$$D_b(\beta_\ell, \alpha_\ell) \leq M I(q_1^{(\ell)}, q_0^{(\ell)}) \tag{5.24}$$

where $q_0^{(\ell)}$ and $q_1^{(\ell)}$ are the densities of sample $x_\ell(n)$ under hypotheses H_0 and H_1, respectively, and

$$I(q_1^{(\ell)}, q_0^{(\ell)}) = \int \log \frac{q_1^{(\ell)}(x)}{q_0^{(\ell)}(x)} q_1^{(\ell)}(x) dx = \frac{s_\ell^2}{2\sigma_\ell^2} \tag{5.25}$$

is the Kullback-Leibler divergence. Combining (5.23)–(5.25), we have

$$\eta_{dis} \leq \frac{1}{M} \sum_{\ell=1}^{L} \max_{h_\ell} M \frac{s_\ell^2}{2\sigma_\ell^2} = \sum_{\ell=1}^{L} \frac{s_\ell^2}{2\sigma_\ell^2}$$

and thus $\eta_{non} \geq \eta_{dis}$.

5.6 Examples of the Performance Computations

In this section, we illustrate the performance of procedures $\mathcal{P}_0$-$\mathcal{P}_3$ for a distributed detection system with $L = 4$ sensors. The sample noise at each sensor is standard Gaussian. Two cases will be considered:

	post-disorder signal strength, dB			
case	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
"strong signal"	0	−3	−6	−10
"weak signal"	−20	−23	−26	−30

For the distributed procedures, the thresholds for the local detectors are obtained using the method of Section 4.

5.6.1 Procedures Where the Jump Magnitudes Are Known

In Figure 5.5, a plot of D vs. T for the ML optimal procedure is shown for the blocklength $M = 20$ and several values of μ. This illustrates the effect of the disorder arrival time on the expected delay. The *average* and *worst-case* performance is also shown. To compute the former, the average expected delay over all μ for each T is determined (i.e., assign a uniform prior probability to μ); for the latter, the maximum expected delay over μ is determined. We observe that D is lower when the disorder occurs either very early or very late in the block. Since μ is unknown, only the average and worst-case performance will be considered from here on.

The accuracy of all of the procedures was verified by performing Monte Carlo simulations. [9] Each of the procedures was implemented on a MasPar massively parallel computer, enabling us to perform multiple runs simultaneously; in this case, each of the 4096 processors executed a single run. In Figure 5.5, the circles are the values

[9]The simulated values are shown explicitly only in Figure 5.5. These values are not shown in the other graphs so that the detail remains clear where there are several plots on one graph.

generated for $h = 2, 4, 6, \ldots, 14$ in this manner. One can see the excellent agreement between the simulated and computed performance throughout. The asterisks in Figure 5.5 show the computed values for $h = 16, 18,$ and 20; simulated values are not shown for these points due to the large computing time required to generate false alarms in this range.

The average and worst-case performance of the ML optimal, Page, and non-distributed procedures are compared for the strong signal case in Figure 5.6. The performance of the non-distributed procedure is uniformly better than that of the other procedures, as expected; in addition, the advantage of utilizing all of the sensor data at the central processor increases linearly in the log of T. It is interesting to note that both the average and worst-case performances of the ML optimal and Page procedures are virtually identical, with the ML optimal winning out only by a small amount in some places. We also see that the performance degrades with increasing blocksize, and the best choice is $M = 1$. This reflects the tradeoff between using large enough M so that the local detectors are sufficiently powerful and keeping M small so that the detection will be quick. In this case, the disorder has a large enough magnitude that only one sample is required for the local detectors.

The same computations are shown for the weak signal case in Figure 5.7. Again, the non-distributed procedure is the best of the three, and the average and worst-case performance for the ML optimal and Page procedures are nearly the same. In general, the expected delays are much larger for the small-signal case; this is also expected, since weaker signals will lead to local detectors with lower power. A major difference between the weak and strong signal cases is the sensitivity of the performance with respect to blocksize. In Figure 5.6, we see that the choice of $M = 10$ yields an expected delay that differs from $M = 1$ by about 5 samples, while for $M = 20$, this difference is closer to 10 samples, a significant percentage of the delay for $M = 1$. However, for the case shown in Figure 5.7, the performance for $M = 500$ is very

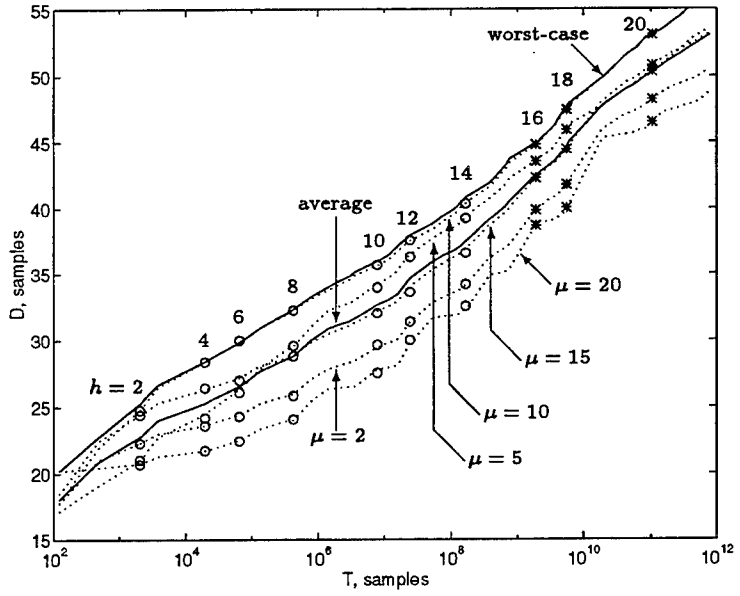

Figure 5.5: Performance of the ML optimal procedure for different μ. $L = 4$, $M = 20$, strong signal case. Circles = simulated values, asterisks = computed values.

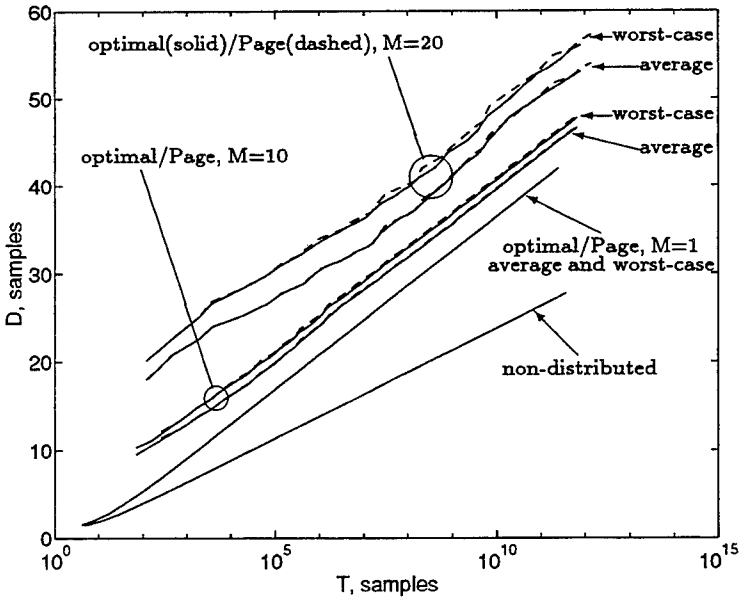

Figure 5.6: Average and worst case performance of the ML optimal, Page, and non-distributed procedures. $L = 4$, strong signal case.

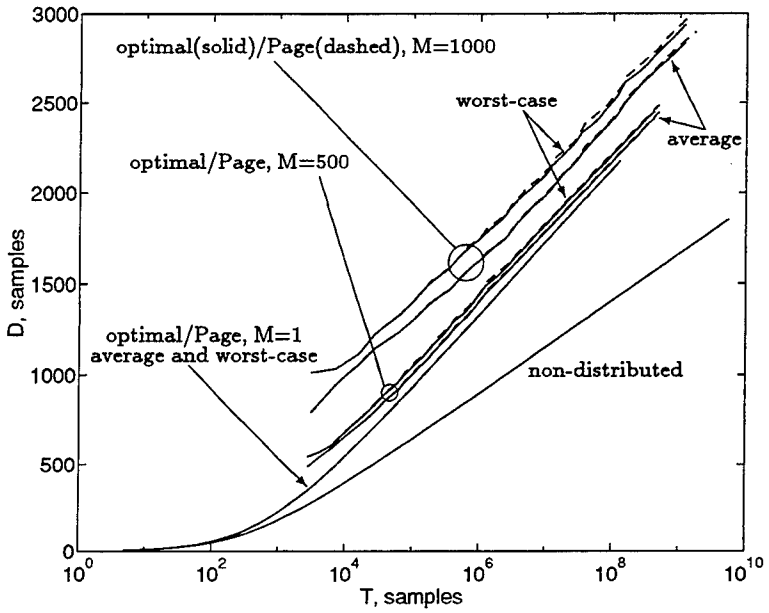

Figure 5.7: Average and worst case performance of ML optimal, Page, and non-distributed procedures. $L = 4$, weak signal case.

close to that for $M = 1$. In addition, the difference in expected delay for these two cases diminishes as T increases. Therefore, the smaller the jump magnitude, the less critical the choice of the blocksize is, though in either case an M which is too large will result in significant performance degradation.

5.6.2 Procedures Where the Jump Magnitudes Are Unknown

We now compute the performance of Hinkley's test and compare it to that of the ML optimal and non-distributed tests (i.e., procedures where the jump magnitudes are known). The jump magnitudes are those of the strong signal case given in the previous section. For Hinkley's test, we consider two magnitudes for the minimum jump: $\text{SNR}_{min} = -10$ and -20 dB. In other words, we determine the parameters

for Page's tests assuming that the disorder is of magnitude SNR_{min}.

The results are shown in Figure 5.8. Here, the average performance is computed (the results are similar when the worst-case performance is used). One can see immediately the benefit of knowing the actual signal strength. The performance of the Hinkley procedures diverges from that of the ML optimal procedure as T increases, regardless of the choice of M. Also, the performance for $\text{SNR}_{min} = -10$ dB exceeds that of $\text{SNR}_{min} = -20$ dB, more so with larger T. This reflects the fact that the better an idea one has about the jump magnitude, the better the performance that can be achieved. Although not included here, a similar comparison was done for the weak-signal scenario, and the results were analogous to the above.

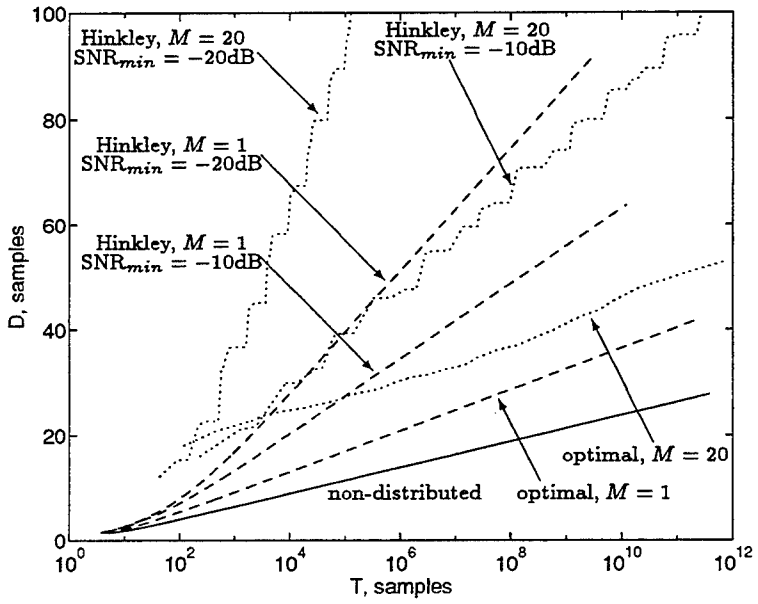

Figure 5.8: Average performance of Hinkley procedure with $\text{SNR}_{min} = -10$ and -20 dB, the ML optimal procedure, and the non-distributed procedure. $L = 4$, strong signal case.

5.7 Sensitivity of Performance To Variation in the Local Thresholds

In Section 4.4, a procedure for determining the thresholds of the local detectors based upon the lower bound of the asymptotic performance measure $\widetilde{\eta}$ was presented. However, it is not clear as to whether the resulting choice of thresholds is near the optimal for two reasons: (1) η is an *asymptotic* performance measure, so the performance in general may be inadequate, and (2) the tightness of the lower bound was not considered. This motivates us to perform the following sensitivity analysis.

Let $\mathbf{h}_{loc}$ denote the L–dimensional vector of local thresholds obtained via the method of Section 4.4, and let $\mathbf{h}_{per}$ denote a multiplicative perturbation of $\mathbf{h}_{loc}$ such that

$$\mathbf{h}_{per} = c \cdot \mathbf{h}_{loc}, \quad c \in \Re_+$$

Thus, by varying the perturbation parameter c and computing the performance using $\mathbf{h}_{per}$ as the local thresholds, we can determine whether $\mathbf{h}_{loc}$ (i.e., $c = 1$) is a good choice.

In Figure 5.9, we compute the average ML optimal performance for $N = 10$ and values of c ranging from 0.5 to 1.5. We see that asymptotically as T increases, $c = 1$ is the best choice. We also note that a change of plus or minus ten percent does not significantly affect the performance. From this, we conclude that the methodology of Section 4.4 is reasonable.

In Figure 5.10, the analysis is repeated for the weak signal case where $N = 100$. Again, we see that the best performance occurs when $c = 1$. However, there is little difference in performance even for variations in $\mathbf{h}_{loc}$ of as much as fifty percent. In other words, the performance is far less sensitive to the choice of local thresholds when the jump magnitudes are smaller.

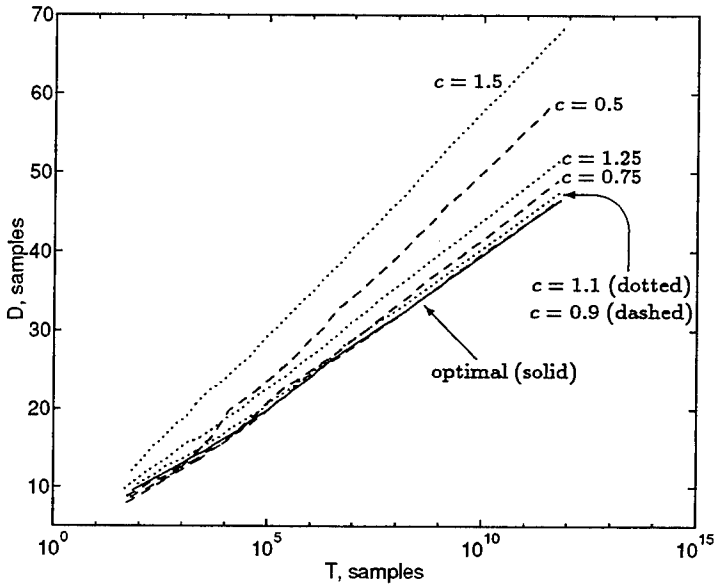

Figure 5.9: Sensitivity of average performance of the ML optimal procedure to perturbations of the local thresholds. $M = 10$, strong signal case.

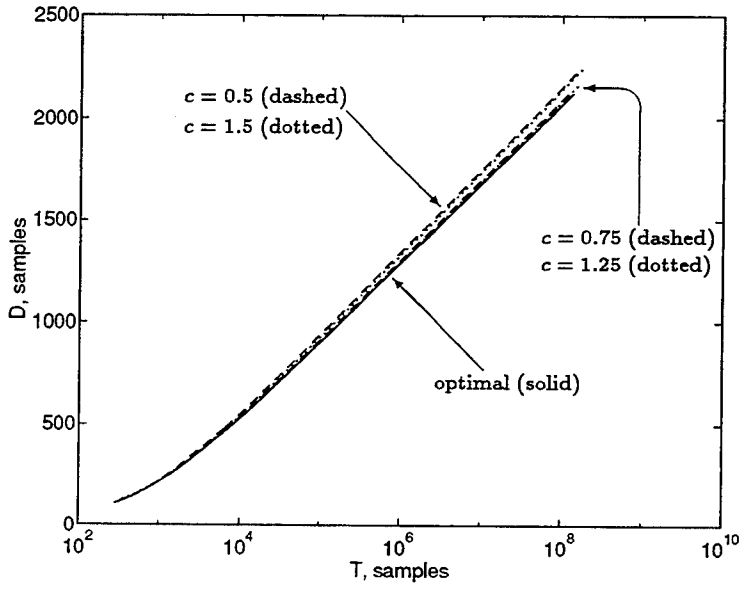

Figure 5.10: Sensitivity of average performance of the ML optimal procedure to perturbations of the local thresholds. $M = 100$, weak signal case.

5.8 Block Length Effects

In this section, we address the question: For a particular distributed system, what is the best choice for the blocksize M? We assume that the number of sensors L is fixed and known, and that the designer has a specified minimum desired T.

There are two considerations in the choice of M. The first is the expected delay for the chosen value of T; in other words, the expected delay can be parameterized as $D(T, M)$. We have seen in the previous section that the performance varies significantly depending on the blocksize.

The second issue is the cost $C(M)$ associated with M. Since the sampling rate f_s is fixed, the smaller the value of M, the more frequently the local decisions are sent to the fusion center, and thus the higher the bandwidth required for each of the channels. Each local decision is represented by a single bit, and so the cost is proportional to the bit rate. Thus:

$$C(M) = C_0 \frac{f_s}{M} \tag{5.26}$$

where C_0 is a constant with units "cost per unit bit rate" which reflects the relative cost of increasing the channel bandwidth. For example, a system with $M = 1$ costs twice as much as a system with $M = 2$, which costs twice again as much as with $M = 4$; i.e. the cost is linearly proportional to the inverse of the blocksize. Note, however, that one is not restricted to costs of the form (5.26); on the contrary, a $C(M)$ which more accurately models the cost structure for a particular system may be substituted. Nonetheless, (5.26) is used in this case without loss of generality.

The best choice of M is a compromise between the desire to quickly detect a disorder and the need to minimize the system cost. To this end, we propose the following design methodology. For procedure $\mathcal{P}_i$, $i = 1, 2, 3$, define the function

$$\mathcal{R}_i(T^*, M) \triangleq \min_{T \geq T^*} \frac{D_i(T, M)}{D_0(T^*)} - 1$$

$\mathcal{R}_i(T^*, M)$ is a *relative performance measure* which reflects the *sacrifice* in performance for using procedure $\mathcal{P}_i$ instead of the non-distributed procedure for the specified minimum allowable mean time between false alarms T^* and block size M. The plot of D vs. $\log T$ for the non-distributed case is continuous; therefore, any operating point T^* is achievable using $\mathcal{P}_0$. However, the fact that the set of possible local decisions at any time is a discrete finite set (of cardinality 2^L) means that a designer may not be able to design a test operating at exactly T^*. [10] Therefore, the delay corresponding to the smallest $T \geq T^*$ is used.

The overall desirability of different block lengths is determined by evaluating both the relative performance function and the cost function over a range of M, and plotting $\mathcal{R}_i(T^*, M)$ versus $C(M)$ for each procedure $\mathcal{P}_i, i = 1, 2, 3$. This allows a system designer to decide whether an incremental cost increase will produce a worthwhile improvement in performance.

In Figure 5.11, the above methodology is illustrated for the strong signal scenario considered previously. Here $\mathcal{R}_i(T^*, M)$ is plotted versus $C(M)$ for $i = 1, 2$, and 3, where for simplicity we take $C_0 = 1$ and $f_s = 1$. The plots are generated for $T^* = 10^6$ and $T^* = 10^{10}$. We see that in general, the relative performance deteriorates as the blocklength M increases; that is, as the system cost decreases. For the ML optimal procedure, the best performance is achieved for $M = 1$; of course, this is also the most expensive alternative. However, a designer might be willing to sacrifice some performance in order to reduce the overall cost. In such cases, the plot of relative performance versus cost makes it easier to select a compromise. For example, one can see that little is sacrificed by using $M = 10$ rather than $M = 1$, and so this is an attractive alternative. The aforementioned fact that the ML optimal ($i = 1$) and Page ($i = 2$) procedures exhibit nearly identical performance is evident here as well.

In Figure 5.12, the relative performance is shown for the weak signal case. Again

[10] One might consider using a randomized test to exactly achieve T^*.

we see that for sufficiently large M, the performance rapidly decreases. However, unlike in the strong signal case, the best choice of blocksize does not correspond to the system with the highest cost. In fact, the best choice for M is closer to 100, and significant performance degradation doesn't occur until M is larger than 500. Another observation is that at $M = 100$, it appears that the performance of the Page procedure is better than that of the ML optimal, yet this is not so. This appearance is a consequence of the fact that we cannot design a test for *exactly* an arbitrary T, as we mentioned earlier; in this instance, the smallest $T \geq T^*$ for the Page procedure was closer to T^* than the corresponding T for the ML optimal procedure, which led to $\mathcal{R}_1(T^*, M) > \mathcal{R}_2(T^*, M)$.

One question arises from the above observations: Why is it better to use larger blocksizes when detecting small signals, even though the goal is quickest detection? The optimal procedure for detecting one mean versus another in Gaussian noise given M observables is to compare the sum of the samples to a threshold. If the mean is large, it will require relatively few samples in order to get a test of reasonable power. For the strong signal case we considered, $M = 1$ was sufficiently large to enable local tests to make decisions with high accuracy. Although a higher M would have produced even more powerful local tests, this increase is more than offset by the additional delay in detection incurred. For weaker signals, M must be larger in order to produce local tests with sufficiently high power. To clarify this point, suppose that $M = 1$ is chosen. The overall procedure then amounts to hard-limiting the sensor measurements and using the outcomes in a sequential procedure at the fusion center, a well-known non-parametric procedure. The fact that $M = 1$ is the best choice in Figure 5.11 shows that for strong signals, it is not necessary to process the local data in an optimal fashion (i.e., summing the samples) in order to get good performance. On the other hand, Figure 5.12 illustrates that for weaker signals, it is necessary to involve many samples in each local decision in order to achieve reasonable power.

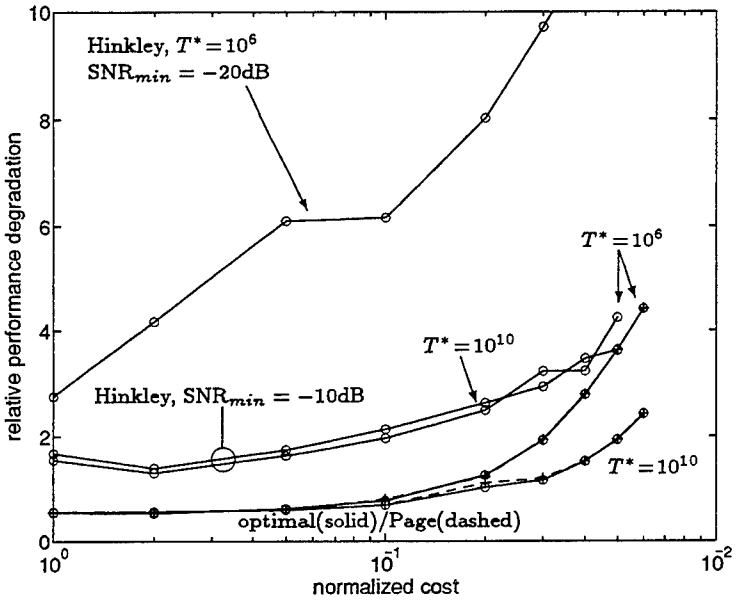

Figure 5.11: $\mathcal{R}_i(T^*, M)$ versus $C(M)$ for the strong signal case.

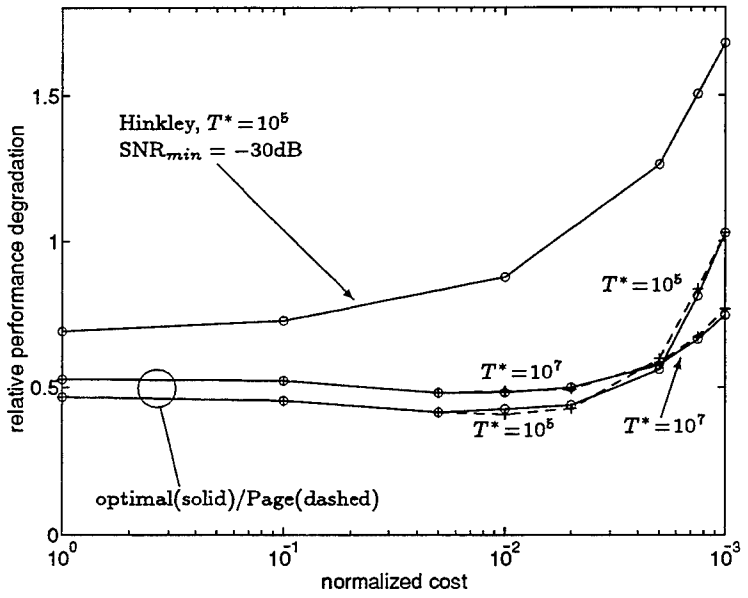

Figure 5.12: $\mathcal{R}_i(T^*, M)$ versus $C(M)$ for the weak signal case.

This tradeoff has been explored previously in the context of quickest detection for scalar (non-distributed) signals [12].

5.9 Conclusions

In this chapter, we examined the quickest distributed detection problem. The local detection procedure was to simply compare the sum of successive blocks of sensor data to a threshold, and the disorder time was assumed unknown. Several alternatives for the fusion procedure were considered. For the case when the magnitudes of the jumps are known, we derived the ML optimal procedure along with a suboptimal version which requires less computation to implement. When the magnitudes are unknown, a similar procedure designed to react to a nominal jump is used. Each of these procedures were shown to have recursive implementations.

The performance of the above procedures was computed by modelling the test statistics as a Markov process, allowing us to get explicit expressions for the average sample numbers before and after the disorder. Analytical expressions for the performance are important since they allow the system designer access to alternatives without extensive Monte Carlo simulations; in this case, the analytic computations were verified via simulation. It was shown that the performances of the ML optimal and suboptimal procedures are asymptotically equivalent, and so in practice the latter test might be the better alternative.

A simple method for choosing the thresholds of the local detectors based upon a lower bound on the asymptotic performance measure was introduced. Sensitivity analysis reveals that the procedure leads to near optimal performance. This eliminates the necessity of solving a set of coupled nonlinear equations to obtain the optimal threshold settings, which is the common situation in decentralized detection problems.

Finally, a methodology was developed to determine the best choice of blocklength

for the local tests. This involved a tradeoff between decision delay and increased system bandwidth (i.e., communication cost). For the strong signal case, the results were as expected: a decrease in communication cost results in a deterioration in performance. Surprisingly, for the weak signal case, there is a range where lowering the communication cost results in an improvement in performance.

There are several interesting directions for future work. First, although we used simple summing devices at the local detectors, the analysis could be easily extended to include other types of detection schemes. For example, if the signal were time varying, two options might be viable: for coherent detection, an estimator-correlator could be used, while for noncoherent detection, a generalized energy detector might be appropriate. Another issue is the assumption of the independence of the samples. If the samples are correlated, then so will be the local decisions. For this case, one approach would be to design Page's test using the conditional densities (conditioned on the past decisions) rather than the marginals; however, if the blocksize is large, the decisions may only be slightly correlated, and so such a modification might not be necessary. Finally, it would be useful to develop a procedure for estimating the *location* of a disorder using a distributed detection scheme. This could be useful, for instance, in the case where distributed sensors monitor seismic activity, and upon detecting an earthquake, one wishes to locate the epicenter.

5.10 Appendices

5.10.A Extension to the Continuous Time Case

Here we present the continuous time analogue to the distributed detection problem. The observables at each sensor are typically modelled using a stochastic differential equation [4, 19]:

$$dx_\ell(t) = s_\ell \cdot u(t - t_0)dt + \xi_\ell \cdot dw_\ell(t),\ \ell = 1, 2, \ldots, L$$

where $w_\ell(t)$ is a Wiener process at sensor ℓ scaled by ξ_ℓ, $u(\cdot)$ is the unit step function, t_0 is the disorder time, and s_ℓ is the drift at sensor ℓ resulting from the disorder.

In the discrete time case, the local decisions are produced by comparing sums of N successive samples to a threshold. In the present case, the summation is replaced by the integral of the sensor measurements over a time window of length NT_s. In particular, $u_\ell(k)$ (the kth decision at sensor ℓ) is

$$u_\ell(k) = \begin{cases} 1, & w_\ell(k) > h_\ell \\ 0, & \text{otherwise} \end{cases}$$

where

$$w_\ell(k) = \int_{(k-1)NT_s}^{kNT_s} x_\ell(t)dt$$

Let $\xi_\ell = \frac{\sigma_\ell}{\sqrt{T_s}}$, and let k_0 and μ denote the change block and the disorder time within block k_0 as before. In the continuous case, the disorder may occur at any time, not just at discrete sample instants; thus $\mu = k_0 NT_s - t_0 \in [(k_0 - 1)NT_s, k_0 NT_s]$. Now the distribution of the local decisions is exactly that of (5.5)-(5.6), with the exception that μ now takes on continuous rather than discrete values. This means that all of the techniques developed for the discrete time case can also be implemented in continuous time.

5.10.B Conversion From Blocks To Samples

Let $\widetilde{N}$ denote the stopping time in terms of blocks, respectively, where a block consists of M samples. Let $p_i(k)$ denote the probability that the alarm occurs at block k under H_i, and let μ denote the number of samples taken from H_1 present in the change block k_0. [11] Since global decisions are only made at the ends of a block (i.e.- every M samples), under H_0 we have

$$\begin{aligned}\mathcal{N}_0(\theta_0) &= Mp_0(1) + 2Mp_0(2) + 3Mp_0(3) + \dots \\ &= M\sum_{k=1}^{\infty} kp_0(k) \\ &= M \cdot \mathrm{E}_0\widetilde{N}\end{aligned}$$

Under H_1, detecting the delay at block k_0 corresponds to a delay of μ samples. For each additional block before the alarm, M additional samples are required. Therefore:

$$\begin{aligned}\mathcal{N}_0(\theta_1) &= \mu p_1(1) + (\mu + M)p_1(2) + (\mu + 2M)p_1(3) + \dots \\ &= \sum_{k=1}^{\infty} [\mu + M(k-1)]\, p_1(k) \\ &= \mathrm{E}_1\left[\mu + M(\widetilde{N} - 1) \mid \mu\right] \\ &= \mu + M\left(\mathrm{E}_1[\widetilde{N} \mid \mu] - 1\right)\end{aligned}$$

where the last equality results from the linearity of the conditional expectation.

[11] Notice that $p_1(k)$ is a function of of μ.

5.10.C Derivation of the Optimal Non-Distributed Procedure

Here we derive the optimal test when all of the data is available at the central processor. For convenience, we adopt the following vector notation:

$$\begin{aligned}
\mathbf{s} &= [s_1, s_2, \ldots, s_L]^T \\
\mathbf{n}(i) &= [n_1(i), n_2(i), \ldots, n_L(i)]^T \\
\mathbf{x}(i) &= \begin{cases} \mathbf{n}(i), & i = 1, 2, \ldots, m-1 \\ \mathbf{n}(i) + \mathbf{s}, & i = m, m+1, \ldots \end{cases}
\end{aligned}$$

where m is the change time, and $\mathbf{x}(i)$ is a snapshot of the sensor data at sample time i. Let the distribution of $\mathbf{n}(i)$ be

$$p(\mathbf{n}(i)) = \left(\frac{1}{2\pi}\right)^{L/2} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{n}^T(i)\Sigma^{-1}\mathbf{n}(i)\right), \quad \forall i \tag{C.1}$$

where $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_L^2)$. The log-likelihood ratio for a disorder occurring at sample m within a block of n samples is

$$\begin{aligned}
\ell_n(\{\mathbf{x}(i)\}_{i=1}^n; m) &= \log \frac{\prod_{i=1}^{m-1} p(\mathbf{x}(i)) \prod_{j=m}^{n} p(\mathbf{x}(j) - \mathbf{s})}{\prod_{i=1}^{n} p(\mathbf{x}(i))} \\
&= \sum_{j=m}^{n} \log \frac{p(\mathbf{x}(j) - \mathbf{s})}{p(\mathbf{x}(j))}
\end{aligned} \tag{C.2}$$

Substituting (C.1) into (C.2), we have

$$\begin{aligned}
\ell_n(\{\mathbf{x}(i)\}_{i=1}^n; m) &= \sum_{j=m}^{n} \left\{ \mathbf{s}^T \Sigma^{-1} \mathbf{x}(i) - \frac{1}{2}\mathbf{s}^T \Sigma^{-1} \mathbf{s} \right\} \\
&= \sum_{j=m}^{n} \sum_{\ell=1}^{L} \sigma_\ell^{-2} \left\{ x_\ell(j) s_\ell - \frac{1}{2} s_\ell^2 \right\} \\
&= \sum_{j=m}^{n} z_j
\end{aligned} \tag{C.3}$$

where the definition of z_j is obvious. The likelihood ratio is now maximized over all possible change times, resulting in the test:

$$S_n = \max_{1 \le m \le n} \sum_{j=m}^{n} z_j \tag{C.4}$$

$$N = \inf\{n \mid S_n > h\} \tag{C.5}$$

This is simply the univariate Page's test, and it is shown in [11] that the recursive test

$$\begin{aligned} S_n &= \max\{S_{n-1} + z_n, 0\} \\ N &= \inf\{n \mid S_n > h\} \end{aligned}$$

is equivalent to (C.4)-(C.5).

The asymptotic performance measure can now be computed. Since the log-likelihood ratio is used to process the data, the lower bound $\tilde{\eta} \leq \eta$ is tight [4]. Therefore,

$$\eta = \tilde{\eta} = \mathrm{E}_1 z_j = \sum_{\ell=1}^{L} \frac{s_\ell}{2\sigma_\ell^2}$$

Therefore, the asymptotic performance is proportional to the collective signal to noise ratios at each of the sensors.

5.10.D Computation of $\tilde{\eta}$ for the Distributed Procedures

Here, it is shown that $\tilde{\eta}$ is a valid lower bound on the asymptotic performance measure η for the distributed detection case. Throughout, $\widetilde{N}$ and N denote stopping times in terms of blocks and samples, respectively.

Consider the following hypothesis testing problem. Let $\mathbf{u}(n) = [u_1(n), \ldots, u_L(n)] \in \{0,1\}^L$ denote the snapshot of local decisions for block n such that

$$\Pr\{u_\ell(n) = 1\} = \begin{cases} \alpha_\ell, & \text{under hypothesis } K_0 \\ \beta_\ell, & \text{under hypothesis } K_1 \end{cases}, \quad \ell = 1, 2, \ldots, L \tag{D.1}$$

where at the disorder time the transition $K_0 \mapsto K_1$ occurs. Define the function $g : \{0,1\}^L \to \Re$. Page's procedure for problem (D.1) is:

$$\mathcal{P}_R : \quad \begin{cases} R_n &= \max\{R_{n-1} + g(\mathbf{u}(n)), 0\}, \quad R_0 = 0 \\ \widetilde{N}_R &= \inf\{n \mid R_n > h\} \end{cases}$$

This test is optimal when g is the log-likelihood ratio for testing K_0 vs. K_1. One may recognize that problem (D.1) is the same as (5.6) for the case when the disorder occurs at the beginning of a block, i.e. $\mu = M$ (so $K'_\mu = K_1$). In [4], the following bounds on ABN are derived:

$$\mathrm{E}_1\widetilde{N}_R \;\leq\; \frac{h+\gamma}{\mathrm{E}_1\left[g(\mathbf{u})\right]} \tag{D.2}$$

$$\mathrm{E}_0\widetilde{N}_R \;\geq\; \exp\{h\omega\} \tag{D.3}$$

where $\gamma > 0$ is a finite constant and ω_0 satisfies the moment generating function identity under H_0, $\mathrm{E}_0[\exp\{w_0 g(\mathbf{u})\}] = 1$. Since g is the log-likelihood ratio, $\omega_0 = 1$. [12] Since, here, the disorder is assumed to occur only at the beginning of a block ($\mu = M$), the bounds on ASN are simply given by (D.2) - (D.3) multiplied by M. Of course, we are actually interested in the performance for arbitrary μ.

We now derive ASN bounds that are applicable for any $\mu \in \{1, 2, \ldots, M\}$ (i.e. the problem in (5.6)) for procedures of the form

$$\mathcal{P}_S: \quad \begin{cases} S_n &= \max\{S_{n-1} + g(\mathbf{u}(n)), \psi(\mathbf{u}(n))\}, \quad S_0 = 0 \\ \widetilde{N}_S &= \inf\{n \mid S_n > h\} \end{cases}$$

where $\psi \in \Re_+$ is bounded. $\mathcal{P}_S$ is a generic procedure which includes the ML optimal procedure $\mathcal{P}_1$ and the suboptimal procedure $\mathcal{P}_2$ (the latter, by letting $\psi \equiv 0$). Let T_S and D_S denote the expected stopping times in samples for $\mathcal{P}_S$ under H_0 and H_1, respectively.

Proposition 5:

$$D_S \leq M\left(1 + \frac{h+\gamma}{E_1[g(\mathbf{u})]}\right) \tag{D.4}$$

[12] Further details regarding γ and ω_0 are given in [4]. However, for the present purposes all necessary information is contained in this appendix.

Proof:

Consider the test

$$\mathcal{P}_U: \quad \begin{cases} U_n &= \begin{cases} 0, & n=1 \\ \max\{U_{n-1}+g(\mathbf{u}(n)),0\}, & n=2,3,\ldots \end{cases} \\ \widetilde{N}_U &= \inf\{n \mid U_n > h\} \end{cases}$$

This procedure differs from $\mathcal{P}_S$ in that: *i.*) the first block of data (stage $n=1$) is neglected, and *ii.*) $\phi(\cdot)$ is replaced by 0. At first, it may be unclear as to why $\mathcal{P}_U$ was chosen. There are two reasons for this choice. First, by ignoring the first block (of which μ samples are from H_1), we have constructed a procedure that is independent of μ; this fact, along with replacing $\phi(\cdot)$ by 0, will enable us to relate the expected stopping time of $\mathcal{P}_S$ to the bound in (D.2). Second, the ABN of $\mathcal{P}_U$ upper bounds that of $\mathcal{P}_S$. To see this, fix a particular realization of $\{\mathbf{u}(n)\}_{n=1}^{\infty}$. By comparing the test statistics termwise, it is clear that $U_n \leq S_n$ for all n, and since this is true for any realization

$$\mathrm{E}_1\widetilde{N}_S \leq \mathrm{E}_1\widetilde{N}_U \tag{D.5}$$

That is, it will take U_n longer to reach the boundary h than it will S_n. Also, since $\mathcal{P}_U$ is independent of μ,

$$\mathrm{E}_1[\widetilde{N}_U \mid \mu = M] = \mathrm{E}_1\widetilde{N}_U \tag{D.6}$$

Now when $\mu = M$, the *expected* sample path of U_n is the same as that of R_n delayed (shifted to the right) by one unit. Thus, the expected stopping time of $\mathcal{P}_U$ satisfies:

$$\mathrm{E}_1[\widetilde{N}_U \mid \mu = M] = 1 + \mathrm{E}_1[\widetilde{N}_R \mid \mu = M] \tag{D.7}$$

Combining (D.5)-(D.7) and using the upper bound in (D.2), we have

$$\begin{aligned} \mathrm{E}_1\widetilde{N}_S &\leq 1 + \mathrm{E}_1[\widetilde{N}_R \mid \mu = M] \\ &\leq 1 + \frac{h+\gamma}{\mathrm{E}_1[g(\mathbf{u})]} \end{aligned} \tag{D.8}$$

Finally, using (5.4) to convert into units of samples along with (D.8):

$$\begin{aligned} D_S &= \mathcal{N}_0(\theta_1) \\ &= (\mathrm{E}_1\widetilde{N}_S - 1)M + \mu \\ &\leq M \cdot \frac{h+\gamma}{\mathrm{E}_1[g(\mathbf{u})]} + \mu \\ &\leq M\left(1 + \frac{h+\gamma}{\mathrm{E}_1[g(\mathbf{u})]}\right) \end{aligned}$$

∎

Proposition 6: *There exists some finite $\rho \geq 0$ such that for $h > \rho$:* [13]

$$T_S \geq M \exp\{(h-\rho)\omega\} \tag{D.9}$$

Proof:

Define the procedure

$$\mathcal{P}_V : \quad \begin{cases} V_n &= \max\{V_{n-1} + g(\mathbf{u}(n)), \rho\}, \quad V_0 = \rho \\ \widetilde{N}_V &= \inf\{n \mid V_n > h\} \end{cases}$$

for some $h > \rho$ where

$$\rho \stackrel{\Delta}{=} \max_{\mathbf{u}\in\{0,1\}^L} \psi(\mathbf{u})$$

and $\rho < \infty$ since ψ is bounded. For any fixed sample realization of $\{\mathbf{u}(n)\}_{n=1}^{\infty}$, $S_n \leq V_n$, $\forall n$. Therefore

$$\mathrm{E}_0\widetilde{N}_S \geq \mathrm{E}_0\widetilde{N}_V \tag{D.10}$$

Now consider the procedure

$$\mathcal{P}_W : \quad \begin{cases} W_n &= \max\{W_{n-1} + g(\mathbf{u}(n)), 0\}, \quad W_0 = 0 \\ \widetilde{N}_W &= \inf\{n \mid S_n > h - \rho\} \end{cases}$$

[13] Note that the restriction $h > \rho$ will not pose a problem, since we will eventually be taking the limit as $h \to \infty$.

This is identical to procedure $\mathcal{P}_V$ except that the lower boundary is shifted from ρ to 0, and the upper boundary is shifted from h to $h-\rho$. It is not difficult to see that for a particular sample path $W_n = V_n - \rho, \forall n$. Therefore,

$$\mathrm{E}_0\widetilde{N}_V = \mathrm{E}_0\widetilde{N}_W \tag{D.11}$$

Now procedure $\mathcal{P}_W$ differs from $\mathcal{P}_R$ only in that the threshold is shifted by ρ. Thus, the lower bound (D.3) also applies to the former, so we have

$$\mathrm{E}_0\widetilde{N}_W \geq \exp\{(h-\rho)\omega\} \tag{D.12}$$

Finally, combining (D.10)-(D.12) and applying (5.3), we obtain

$$\begin{aligned} T_S &= \mathcal{N}_0(\theta_0) \\ &= M \cdot \mathrm{E}_0\widetilde{N}_S \\ &\geq M \cdot \mathrm{E}_0\widetilde{N}_W \\ &\geq M\exp\{(h-\rho)\omega\} \end{aligned}$$

∎

By substituting (D.4) and (D.9) into the definition of η, we get the desired result:

$$\begin{aligned} \eta &= \lim_{h\to\infty} \frac{\log T_S}{D_S} \\ &\geq \lim_{h\to\infty} \frac{(h-\rho)\omega + \log M}{\left(1+\frac{h+\gamma}{\mathrm{E}_1[g(\mathbf{u})]}\right)M} \\ &= \frac{1}{M}\omega\mathrm{E}_1\left[g(\mathbf{u})\right] = \widetilde{\eta} \end{aligned}$$

5.11 References

[1] M. Basseville, "On-line detection of jumps in the mean," in *Detection of Abrupt Changes in Signals and Dynamical Systems*, M. Basseville and A. Benveniste, eds., Springer-Verlag, New York, pp. 11-26, 1986.

[2] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[3] R. E. Blahut, *Principles and Practices of Information Theory*, Addison-Wesley Publishing Co., Reading, MA., 1987.

[4] B. Broder, *Quickest Detection Procedures and Transient Signal Detection*, Ph.D. Thesis, Princeton University, 1990.

[5] Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. Aero. Electron. Sys.*, vol. AES-22, no. 1, Jan. 1986.

[6] R. W. Crow and S. C. Schwartz, "Quickest detection for distributed sensor systems," *Proc. of the 31st Allerton Conf. on Commun., Control, and Computing*, pp. 748-757, 1993.

[7] E. Drakopoulos and C. C. Lee, "Optimum multisensor fusion of correlated local decisions," *IEEE Trans. Aero. Electron. Sys.*, vol. AES-27, no. 4, Jul. 1987.

[8] J. Han, P. K. Varshney, and R. Srinivasan, "Distributed binary integration," *IEEE Trans. Aero. Electron. Sys.*, vol. AES-29, no. 1, Jan. 1993.

[9] H. R. Hashemi and I. B. Rhodes, "Decentralized sequential detection," *IEEE Trans. Inform. Theory*, vol. IT-35, no. 3, May 1989.

[10] M. Kam, Q. Zhu, and W. S. Gray, "Optimal data fusion of correlated local decisions in multiple sensor detection systems," *IEEE Trans. Aero. Electron. Sys.*, vol. AES-28, no. 3, Jul. 1992.

[11] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100-114, 1954.

[12] L. Pelkowitz and S. C. Schwartz, "Asymptotically optimum sample size for quickest detection," *IEEE Trans. Aero. Electron. Sys.*, vol. AES-23, no. 2, March 1987.

[13] A. R. Reibman and L. W. Nolte, "Optimal detection and performance of distributed sensor systems," *IEEE Trans. Aero. Electron. Sys.*, vol. AES-23, no. 1, Jan. 1987.

[14] D. Teneketzis, "The decentralized quickest detection problem," *Proc. 21st IEEE Conf. on Decision and Contr.*, Orlando, FL, Dec. 1982, pp. 673-679.

[15] D. Teneketzis and P. Varaiya, "The decentralized quickest detection problem," *IEEE Trans. Auto. Control*, vol. AC-29, no. 7, Jul. 1984.

[16] R. R. Tenney and N. R. Sandell, Jr., "Detection with distributed sensors," *IEEE Trans. Aero. Electron. Sys.*, vol. AES-17, no. 4, Jul. 1

[17] J. N. Tsitsiklis, "On threshold rules in decentralized detection," *Proc. 25st IEEE Conf. on Decision and Contr.*, Athens, Greece, Dec. 1986, pp. 232-236.

[18] V. V. Veeravalli, T. Basar, and H. V. Poor, "Decentralized sequential detection with a fusion center performing the sequential test," *IEEE Trans. Inform. Theory*, vol. IT-39, no. 2, Mar. 1993.

[19] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1985.

Chapter 6

An Adaptive Procedure for Quickest Detection

6.1 Introduction

In Chapter 2, the problem of detecting a shift in the mean of a sequence of independent random variables from θ_0 to θ_1 was considered. There, H_0 and H_1 were defined as the "noise only" and "signal plus noise" hypotheses, respectively. It was shown that, when both θ_0 and θ_1 are known, the optimal procedure based on Lorden's criterion is Page's test implemented using the log-likelihood ratio.

While the assumption of known θ_0 and θ_1 is convenient for simplifying the problem, in many applications one or both parameters may not be known exactly. In this chapter, procedures for detecting a shift of unknown magnitude in the mean of a random process are investigated. It is assumed throughout that θ_0 is known, but that θ_1 is unknown. This is a reasonable assumption, since it is often the case that the state of a system is known or can be adequately estimated before the disorder occurs. For example, in radar applications, the return under the ambient noise hypothesis (that is, when no target is present) may be well-modelled as zero mean white Gaussian noise. However, when a target does appear, the strength of the return, which induces a

proportional change in the mean of the observables, is dependent on several variables; these include the distance and size of the target, as well as propagation effects such as scattering and multipath. In such an instance, if an incorrect value for θ_1 were used to model the system, the actual performance could deviate greatly from that which was computed based on the assumed parameter values.

The actual distribution of the random variables depends on the particular application. We choose to focus on the case of Poisson observables, although the basic techniques can be used for other distributions as well. Disorder problems with Poisson observables have potential applications in many areas. For example, various medical imaging techniques involve the generation of a picture whose pixel intensities are proportional to the number of photons incident on the detector. In many cases, the safety of the patient necessitates keeping the radiation dose at a minimum, resulting in relatively low photon counts; in this case, the Gaussian assumption often used in image processing may be invalid. [1] A sequential detection scheme which incorporates the Poisson assumption directly could be used in the line-by-line detection of boundaries in the image. There are also queueing system applications: one might wish to detect changes in highway traffic flow or in packet arrival times at a server, both of which are often modelled using the Poisson distribution. Finally, in optical communications, the variation in the arrival rate of photons at the receiver could be monitored.

The problem is stated precisely in Section 2. Next, in Section 3, we review some of the commonly used approaches for detecting jumps of unknown magnitude in the mean and discuss the advantages and disadvantages of each. In Section 4, we introduce a new adaptive procedure for detecting such a change with the following properties: *i)* the procedure is recursive, making it useful where an on-line algorithm is

[1] In other words, there may be an insufficient number of samples for the Central Limit Theorem (leading to the Gaussian assumption) to be applicable.

desired, and *ii)* the performance is similar to that of the optimal Page's test when the true θ_1 is known. This procedure consists of two independent stages which operate sequentially. The first stage is a version of Page's test that is useful in detecting jumps of at least some minimum known magnitude. The second stage is an adaptive version of the classical sequential probability ratio test which incorporates an on-line estimate of the mean of the observables. In Section 5, the performance of the adaptive procedure is analyzed and computed for several examples. Due to the difficulty of obtaining closed-form analytical expressions, most of the results in this chapter are based on Monte Carlo simulations. [2]

6.2 Problem Statement

Let $f(x \mid \theta)$ denote a density function with mean θ. Let $X_1, X_2, \ldots$ denote a sequence of random variables generated under the following hypotheses:

$$H_0: \quad X_i \sim f_0(x) \triangleq f(x \mid \theta_0), \qquad i = 1, 2, \ldots, m-1$$

$$H_1: \quad X_i \sim f_1(x) \triangleq f(x \mid \theta_1), \quad \theta_1 \in \Theta \quad i = m, m+1, \ldots$$

Here, Θ is the set of all permissible values of θ_1. Throughout, we will take

$$\Theta \triangleq \{\theta \mid \theta \geq \theta_0 + \nu_0\}$$

where $\nu_0 > 0$ is the minimum possible jump in the magnitude of the mean, which is assumed to be known. [3] That is to say, $\theta_1 = \theta_0 + \nu$, where $\nu \geq \nu_0$. Since θ_0 and θ_1 are constants, we see that the system undergoes a one-time shift from $\theta = \theta_0$ to $\theta = \theta_1$ at the *disorder time* m. The goal is to determine the presence of the disorder as quickly

[2] An earlier version of this procedure appears in [5].

[3] In many applications, the disorder can take on a continuum of values. In such cases, ν_0 is chosen to be the minimum change *of interest* to the designer, as opposed to the minimum change which *can occur*.

as possible. In other words, we wish to minimize the expected delay in detection D for a desired mean time between false alarms T. [4]

As discussed in Chapter 2, when both θ_0 and θ_1 are known, the optimal procedure is Page's test implemented using the log-likelihood ratio processor, $g(x) = \log \frac{f_1(x)}{f_0(x)}$. However, it is often the case that one or both of θ_0 and θ_1 are unknown. In the sequel, we investigate the case where pre-disorder mean θ_0 is known, but the jump magnitude ν is unknown. In particular, the focus will be on jumps which occur in the rate parameter of the Poisson distribution. Thus,

$$f_0(x) = \frac{(\theta_0\tau)^x e^{-\theta_0\tau}}{x!}, \; x = 0, 1, \ldots$$

and

$$f_1(x) = \frac{(\theta_1\tau)^x e^{-\theta_1\tau}}{x!}, \; x = 0, 1, \ldots$$

where θ_0 is known and $\theta_1 \in \Theta$. [5] For simplicity, we will let $\tau = 1$ throughout. There are many applications involving the Poisson distribution where quickest detection procedures would be useful, as discussed in the previous section.

6.3 Conventional Procedures

In this section, some established approaches for detecting jump changes of unknown magnitude are outlined. The most direct approach to this problem involves replacing all of the unknown quantities with their respective maximum likelihood (ML) estimates. For the present problem, the unknowns are the disorder time m and the jump magnitude ν. The resulting procedure is called the *generalized likelihood ratio test* (GLRT).

[4] See Chapter 2 for the precise definitions of D and T.

[5] Notice that a Poisson random variable with rate parameter θ has a mean and variance both equal to $\theta\tau$. Therefore, the jump in θ results in a change not only in mean, but also in variance.

At each time instant n, the pair (k, ν) is chosen to maximize the log-likelihood ratio

$$\log\left[\frac{f_1(X_1,\ldots,X_n)}{f_0(X_1,\ldots,X_n)}\right] = \sum_{i=k}^{n} \log \frac{f(x \mid \theta_0+\nu)}{f(x \mid \theta_0)}$$

Define the test statistic

$$S_n = \max_{\theta\in\Theta} \max_{1\le k\le n} \sum_{i=k}^{n} \log \frac{f(x \mid \theta_0+\theta)}{f(x \mid \theta_0)}$$

The procedure is then to declare a disorder at time n in case

$$S_n > \gamma$$

The threshold γ can be set based either on a criteria of minimum false alarm rate or maximum expected delay. The GLRT is also useful when the probability densities have parametric uncertainty. For example, if the noise densities were known to be Gauss-Gauss mixtures

$$f(x) = \frac{1-\varepsilon}{\sqrt{2\pi}\sigma_0} \exp\left\{\frac{-x^2}{2\sigma_0^2}\right\} + \frac{\varepsilon}{\sqrt{2\pi}\sigma_1} \exp\left\{\frac{-x^2}{2\sigma_1^2}\right\}$$

an ML estimate of the contamination factor ε could also be incorporated into S_n in a straightforward manner.

The GLRT is similar to Page's test in the sense that the test statistic S_n is just the likelihood ratio of all samples up to the current time instant. In fact, the optimal Page's test is just a degenerate version of the GLRT where $\Theta = \{\theta_1\}$. Unfortunately, the GLRT has several undesirable properties. First, unlike with Page's test, an exhaustive search must be performed over all $k = 1, 2, \ldots, n$ and all possible $\theta \in \Theta$. As a consequence, the GLRT does not readily admit a recursive implementation. Second, since $m \in \{1, \ldots, n\}$, the search region increases linearly with n. [6] Third, at each time n, all past and current observables $X_1, \ldots, X_n$ must be stored.

[6] In practical implementations, however, the search region would usually be restricted to some finite interval. Another approach might be to employ some "intelligent" processing which only retains those disorder time candidates which are the most probable based on the past data.

Another approach when the jump magnitudes are unknown is to simply use the optimal procedure designed for the minimum jump ν_0. That is, implement Page's test using the processor

$$g(x) = \log \frac{f(x \mid \theta_0 + \nu_0)}{f(x \mid \theta_0)}$$

Recall that Page's test was defined using the test statistic

$$S_n = \max\{S_{n-1} + g(X_n), 0\}$$

where a disorder is declared when $S_n > h$. Here, we refer to this approach as *Hinkley's procedure*, but it is also known as the Page-Hinkley test for the case when the noise is Gaussian (and so $g(x)$ is linear) [3].

In considering the use of Hinkley's test, care must be taken to ensure that the log-likelihood ratio is increasing (monotonic); such is the case with the Gaussian and Poisson distributions. When this is so, observe that

$$0 < \mathrm{E}[g(x) \mid \theta_0 + \nu_0] \leq \mathrm{E}[g(x) \mid \theta_0 + \nu]$$

for any $\nu \geq \nu_0$. Since the performance is proportional to this expectation (cf. the definition of $\tilde{\eta}$ in Chapter 2), we see that the minimum performance is achieved under the minimum jump scenario. Thus, one can design the procedure to guarantee a nominal level of performance via the selection of the threshold h. In cases when the log-likelihood is not monotonic (for example, with the Gauss-Gauss mixture), Hinkley's test might be a poor choice if the true jump magnitude could take on values in a large range. An alternative would then be to use a suitable monotonic nonparametric nonlinearity for $g(x)$, such as the sign detector or dead-zone limiter. The performance of Page's test using these processors with *known* jump is investigated in [4]; a similar analysis could be done when the magnitudes of the jumps are unknown.

Since Hinkley's procedure is a version of Page's test, it is desirable in that it can also be implemented recursively as explained in Chapter 2. Examples of applications

which utilize Hinkley's test are line-by-line edge detection [1], where the intensities on each side of the boundaries are unknown, and the detection of changes in the quality of links in communications networks [6], where deviations in the nominal probability of bit error are monitored.

Another class of detectors involves the computation of the sample derivative of the mean of the sequence of random variables. This is done by computing a weighted difference of a subset of the samples before and after the hypothesized disorder time. This gives rise to so-called *filtered derivatives* detectors.

In [2], two examples of this type of detector are examined: the integrating filter and a "triangular" filter. For the integrating filter, the statistic

$$Z_n = \frac{X_{n+l} - X_{n-l}}{2l}$$

is used to approximate the derivative of the mean. When Z_n exceeds a threshold, a disorder is declared. The statistic

$$Z_n = \frac{(X_{n+1} + \ldots + X_{n+l}) - (X_{n-1} + \ldots + X_{n-l})}{l^2}$$

is used for the triangular filter in a similar manner.

It is shown [2] that for the filtered derivatives algorithms, both T and D are exponential functions of the chosen threshold. However, for Hinkley's test, T is an exponential function of the threshold, while D is a linear function of the threshold. A consequence of this fact is that for large T, the value of D will be much smaller for Hinkley's test than for the filtered derivatives procedures; hence, it is concluded in [2] that Hinkley's test is superior to filtered derivatives procedures. Furthermore, with respect to the GLRT, Hinkley's procedure has the advantage of being recursive. Therefore, the performance of Hinkley's procedure will be included in the analysis to follow.

6.4 An Adaptive Procedure

The discussion of the previous section motivates us to seek out a procedure with the following properties: *i)* the procedure is recursive, making it useful when an on-line algorithm is desired, and *ii)* the performance is similar to that of the optimal Page's test when the true θ_1 is known. In this section, a heuristic procedure is presented with these goals in mind.

In general, for practical detection and estimation problems, it is often the case that at least one of θ_0 and θ_1 are unknown. A common approach then is to use estimates of the parameters, obtained either on-line or via some historical data. For example, suppose that the samples are distributed as either $\mathcal{N}(\theta_0, \sigma^2)$ or $\mathcal{N}(\theta_1, \sigma^2)$ at *all* time instants (i.e. there is no disorder); this is just the classical hypothesis testing problem. When θ_i, $i = 0, 1$, are known, the sample mean provides a sufficient statistic for deciding between the two hypotheses [8]. For the related composite hypothesis testing problem where the means are unknown and $\theta_0 \leq \theta_0' < \theta_1' \leq \theta_1$, this same procedure can also be used, where the performance does not fall below that of the case where $\theta_i = \theta_i'$, $i = 0, 1$.

For the disorder problem, however, this approach is often not feasible for the following reason. Before the disorder time, all of the data is generated according to $f_0(x)$. If m is relatively large, resulting in a long wait before the disorder occurs, many observables will be available to obtain an estimate of θ_0. This is the case, for instance, in radar problems where the presence of a target occurs only after a long period of time. Even if m is small (excluding the degenerate case where $m = 1$), we at least know that *some* of the samples were generated under $f_0(x)$. By comparison, the estimation of θ_1 is more difficult for two reasons: *i)* the disorder time m is unknown, and therefore so is the instant at which the observables are generated from $f_1(x)$, and *ii)* the use of many samples from $f_1(x)$ in order to obtain an accurate estimate for θ_1

is in competition with the desire to make the decision as quickly as possible. As is usually the case with any sequential detection scheme, the longer one is able to wait, the more accurate a decision can be made.

Based on the above discussion, it is clear that it would be unwise to attempt to estimate θ_1 based on all of the past observations. Instead, it would be desirable to separate the pre- and post-disorder observables for the purpose of estimating θ_0 and θ_1. If m were known, then this would be easy to do; of course, this is also not useful since we would then know when the disorder occurred, which is exactly the problem in the first place!

With this difficulty in mind, we now propose an alternative procedure for detecting jumps in the mean of unknown magnitude, maintaining the assumption that the magnitude of the jump is at least some minimum value ν_0. This procedure consists of two separate tests in series, which will be denoted $\mathcal{T}_1$ and $\mathcal{T}_2$. Test $\mathcal{T}_1$ is exactly the Hinkley test introduced in Section 6.3. Test $\mathcal{T}_2$ is a variation on the classical two-sided SPRT of Wald [7].

When using the SPRT to distinguish between two hypothesized means, it is usually assumed that the two means are known a priori. However, here the value of θ_1 is not known, so we instead use the ML estimate based upon the samples $X_{j+1}, \ldots, X_n$, where j is the most recent stopping time of $\mathcal{T}_1$ and n is the present sample time. Thus, test $\mathcal{T}_2$ is similar to an "estimator-correlator" in the sense that an estimate of the true parameter is correlated with the present data via the log-likelihood ratio; therefore, we call the resulting procedure the *estimator-correlator sequential probability ratio test* (ECSPRT). [7] Let $\theta_{min} \triangleq \theta_0 + \nu_0$ denote the minimum possible value of θ_1. The test statistic for the ECSPRT is

$$\tilde{\ell}_n = \sum_{i=1}^{n} \log \frac{f(X_i \mid \tilde{\theta}_n)}{f(X_i \mid \theta_0)} \tag{6.1}$$

[7] This procedure is not an estimator-correlator in the true sense, but the idea is similar.

where the decisions as to the occurrence of a disorder are made as follows:

$$\tilde{\ell}_n \begin{cases} > b, & \text{decide for } H_1 \\ < a, & \text{reject } H_1 \\ \in (a,b), & \text{compute } \tilde{\ell}_{n+1} \end{cases}$$

Here $\tilde{\theta}_n = \max\{\theta_{min}, \hat{\theta}_n\}$ and $\hat{\theta}_n$ is the ML estimate of θ_1 given n observations, which in this case is the sample mean, and (a,b) is called the *continuation region*, where $a < 0 < b$. Note that $\hat{\theta}_n$ can be computed recursively as

$$\hat{\theta}_n = \frac{n-1}{n}\hat{\theta}_{n-1} + \frac{1}{n}X_n, \ \hat{\theta}_1 = X_1$$

Note also that as $n \to \infty$, $\tilde{\theta}_n \xrightarrow{a.s.} \theta_1$ when $\theta = \theta_1$ and $\tilde{\theta}_n \xrightarrow{a.s.} \theta_{min}$ when $\theta = \theta_0$. Thus, for large n, the ECSPRT behaves like the optimal SPRT under H_1, but like the "minimum jump" SPRT under H_0.

The configuration of the overall adaptive detector, termed the *composite detector*, is shown in Figure 6.1. The purpose of test $\mathcal{T}_1$ is to signal the *possibility* of a disorder. If an alarm sounds during this test, the second test is initiated. If test $\mathcal{T}_2$ results in the acceptance of H_1, then the composite test is terminated and a disorder is declared. If H_1 is rejected by $\mathcal{T}_2$ the composite test is "reset" and $\mathcal{T}_1$ is initialized and started once again. Thus, a false alarm occurs in the composite test if and only if it occurs in both $\mathcal{T}_1$ and $\mathcal{T}_2$.

Let $h > 0$ denote the threshold in $\mathcal{T}_1$, and let $a < 0 < b$ denote the lower and upper thresholds, respectively, in $\mathcal{T}_2$. The composite test is then given by the following algorithm:

1. Initialize: $i = 0, \quad S_0 = 0$

2. Hinkley's Test:

   $i = i + 1$

   $S_i = \max\{0, S_{i-1} + g(X_i)\}$

   if $S_i < h$, goto 2

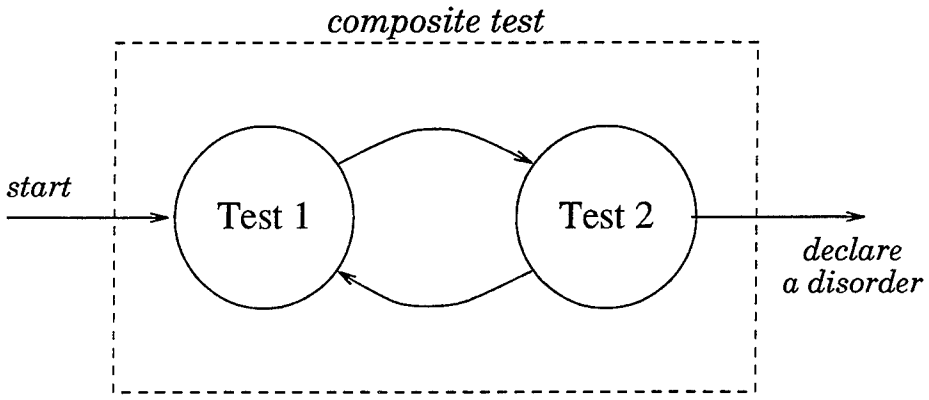

Figure 6.1: Illustration of the composite test.

3. $k = i + 1$

4. ECSPRT:

   $i = i + 1$

   $S_i = \sum_{j=k}^{i} \log \frac{f(X_j|\widetilde{\theta}_i)}{f(X_j|\theta_0)}$

   if $a < S_i < b$, goto 4

   if $S_i \leq a$, set $S_i = 0$ and goto 2

5. Declare a disorder at time i

where $\widetilde{\theta}_i$ is based on samples $X_k, \ldots, X_i$.

In Section 6.3, it was mentioned that when the log-likelihood ratio is not monotonic, the performance of Page's test designed for the minimum jump cannot be guaranteed if the magnitude of the disorder is actually larger. It was suggested that a viable alternative would be to substitute an appropriate memoryless nonlinearity $g(x)$ for the log-likelihood ratio. The same thing can be done in test $\mathcal{T}_1$, which is also just Hinkley's test.

6.5 Performance Evaluation

The performance of the composite test is evaluated by computing D versus $\log T$ for a range of parameters, allowing us to make direct comparisons with the optimal Page's test and with Hinkley's test. An expression for the average sample number (ASN) of the composite test in terms of tests $\mathcal{T}_1$ and $\mathcal{T}_2$ is determined. It is not clear how to obtain explicit expressions for the performance of the ECSPRT, and so we use Monte Carlo simulations to get an estimate of the true performance, as explained below. Several examples that illustrate the performance of the composite test are given.

6.5.1 Analysis

Denote the ASN of $\mathcal{T}_k$ when the rate parameter is θ as $\mathcal{N}_k(\theta)$ for $k = 1, 2$, and let $\mathcal{N}_2^a(\theta)$ and $\mathcal{N}_2^b(\theta)$ be the ASN of $\mathcal{T}_2$ given that the test terminated at the lower and upper boundary, respectively. Tests $\mathcal{T}_1$ and $\mathcal{T}_2$ are independent since the sets of samples that determine their outcomes are disjoint and the samples themselves are independent. Therefore, the two sub-tests can be analyzed separately and these results can then be combined for the composite test.

Denote the ASN of the composite test under θ as $\mathcal{N}(\theta)$, and let

$$\alpha(\theta) = \Pr\{\mathcal{T}_2 \textit{ terminates at the upper boundary } b \mid \theta\}$$

and

$$\beta(\theta) = 1 - \alpha(\theta) = \Pr\{\mathcal{T}_2 \textit{ terminates at the lower boundary } a \mid \theta\}$$

For notational convenience, we temporarily drop the explicit dependence on θ. $\mathcal{N}$ is the sum over j of the expected time to cycle though $\mathcal{T}_1$ and $\mathcal{T}_2$ exactly j times weighted by the probability of exactly j cycles occurring before the test terminates. We have

$$\mathcal{N} = \left(\mathcal{N}_1 + \mathcal{N}_2^b\right) \cdot \alpha + \left(2\mathcal{N}_1 + \mathcal{N}_2^a + \mathcal{N}_2^b\right) \cdot \alpha\beta +$$

$$\begin{aligned}
&\left(3\mathcal{N}_1 + 2\mathcal{N}_2^a + \mathcal{N}_2^b\right) \cdot \alpha\beta^2 + \ldots \\
&= \alpha \cdot \sum_{i=0}^{\infty} \left[\mathcal{N}_1 + \mathcal{N}_2^b + i\left(\mathcal{N}_1 + \mathcal{N}_2^a\right)\right] \beta^i \\
&= \frac{1}{1-\beta}\mathcal{N}_1 + \mathcal{N}_2^b + \frac{\beta}{1-\beta}\mathcal{N}_2^a \\
&= \frac{\mathcal{N}_1 + \mathcal{N}_2}{1-\beta}
\end{aligned}$$

where we use the fact that $\mathcal{N}_2 = \beta\mathcal{N}_2^a + \alpha\mathcal{N}_2^b$. Therefore, the ASN's under each hypothesis can be obtained by computing

$$\mathcal{N}(\theta_i) = \frac{\mathcal{N}_1(\theta_i) + \mathcal{N}_2(\theta_i)}{\alpha(\theta_i)}, \quad i = 0, 1 \tag{6.2}$$

$\mathcal{N}(\theta_0)$ and $\mathcal{N}(\theta_1)$ above are analogous to the T and D of Page's test, respectively. Thus, the two tests can be compared by examining D vs. $\log T$ and $\mathcal{N}(\theta_1)$ vs. $\log \mathcal{N}(\theta_0)$ side by side.

The form of the expression in (6.2) sheds some light on the nature of the composite test. Observe that the ASN is inversely proportional to $\alpha(\theta)$, the probability of crossing the upper threshold in test $\mathcal{T}_2$. When $\theta = \theta_1$, the test statistic in $\mathcal{T}_2$ will have a positive drift, and so the test will terminate at the upper boundary with high probability. Therefore, $\alpha(\theta_1)$ will be relatively close to unity, and so the values of $\mathcal{N}_1(\theta_1)$ and $\mathcal{N}_2(\theta_1)$ will dominate the ASN expression. On the other hand, when $\theta = \theta_0$, the test statistic exhibits a negative drift, resulting in a much smaller $\alpha(\theta_0)$. In this case, $\alpha(\theta_0)$ dominates the ASN.

Unfortunately, it is not clear whether closed form expressions for $\alpha(\theta)$ and $\mathcal{N}_2(\theta)$ can be obtained. Therefore, the approach here is to use Monte Carlo simulations to compute the ASN's. An unbiased estimator of $\mathcal{N}(\theta)$ is

$$\widehat{\mathcal{N}}(\theta) = \frac{1}{K}\sum_{k=1}^{K} N_k \tag{6.3}$$

where a total of K trials are used, and N_k is the stopping time of trial k when the rate parameter is θ. In order to obtain statistically meaningful results, a large number

of trials must be performed for each set of parameters, since the variance of $\widehat{\mathcal{N}}(\theta)$ decreases as $\frac{1}{K}$. One problem with this method, though, is that when $\theta = \theta_0$ the expected stopping time for each trial increases exponentially in the thresholds b and h, and so the time required to complete each trial becomes intractably large.

One way around this problem is to estimate the parameters that appear in (6.2) separately. Observe that the $\mathcal{N}_1(\theta)$ is the ASN of Hinkley's test, a quantity which can be computed using a number of techniques as explained in Chapter 2. Thus, this quantity can be precomputed for the desired threshold h. The remaining parameters, $\mathcal{N}_2(\theta)$ and $\alpha(\theta)$, can be determined by direct Monte Carlo simulation. The advantage of this approach is that only the simulation of the ECSPRT is performed, as opposed to the entire composite test, and therefore the delay associated with the execution of test $\mathcal{T}_1$ is circumvented. The ASN of $\mathcal{T}_2$ can be obtained exactly as in (6.3), while $\alpha(\theta)$ can be approximated by

$$\widehat{\alpha}(\theta) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{I}\left\{S_{N_k}^k > b \mid \theta\right\}$$

where

$$\mathcal{I}\{\mathcal{A}\} = \begin{cases} 1, & \text{if } \mathcal{A} \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

denotes the indicator of event $\mathcal{A}$ and S_i^k is the ECSPRT test statistic for trial k at time sample i. [8]

6.5.2 Comparison of the Performance of Each Procedure

In this section, the composite detector is compared to Hinkley's test and the quickest detector. Although the latter is not realizable in practice since θ_1 would have to be known, it is included as it is the optimal test and therefore provides a useful

[8] It is clear that the frequency of the event $\left\{S_{N_k}^k > b \mid \theta\right\}$ is inversely proportional to the value of b. Thus, it is still possible to select b large enough so that even the alternative method is too computationally burdensome.

standard for performance comparison. Each of the detectors was simulated with Poisson observables with normalized rates $\theta_0 = 10$ and $\theta_{min} = 11$ for several values of θ_1. One can think of θ_0 as the rate under the "noise only" or ambient hypothesis, while $\theta_1 = \theta_{sig} + \theta_0$ is the "signal plus noise," $\theta_{sig} \geq \theta_{min} - \theta_0$. The signal-to-noise ratio (SNR) is $10\log_{10}(\theta_{sig}/\theta_0)$. For example, one can see that the minimum jump of interest here is -10dB.

Realizations of the composite test under hypotheses H_1 (for $\theta_1 = 15$) and H_0 are shown in Figures 6.2 and 6.3, respectively. In the first figure, observe that a disorder is declared only after both of the upper thresholds h (of $\mathcal{T}_1$) and b (of $\mathcal{T}_2$) are crossed. In this example, each of tests $\mathcal{T}_1$ and $\mathcal{T}_2$ are performed once. In Figure 6.3 no disorder occurs and, as one would want, none is indicated (i.e., the threshold b is never crossed in $\mathcal{T}_2$). However, notice that test $\mathcal{T}_1$ signals an alarm two times, but each time test $\mathcal{T}_2$ quickly rejects the supposition that a disorder occurred. This illustrates the "two step" nature of the composite test.

For the composite test, $\mathcal{N}(\theta_1)$ is obtained via direct Monte Carlo simulation, while $\mathcal{N}(\theta_0)$ is obtained by computing $\mathcal{N}_1(\theta_0)$ and simulating the ECSPRT parameters as explained in Section 6.5.1. Figure 6.4 shows a plot of $\widehat{\alpha}(\theta_0)$ versus the upper threshold b for $\mathcal{T}_2$. Observe that $\widehat{\alpha}(\theta_0)$ is an exponential function of b. The smallest value, that corresponding to $b = 13$, was obtained by performing 300,000 trials, which produced 15 false alarms and took several days to run on a Sun 4 workstation.

Figures 6.5-6.7 illustrate the performance of the composite detector as compared to the optimal Page test (the "quickest detector") and Hinkley's test. For the composite test, the plots were generated using Monte Carlo simulations as explained in Section 6.5.1. For the quickest detector and Hinkley's test, the Markov approximation technique discussed in Chapter 2, Section 4 was used.

The comparative performance of the composite detector when $\theta_1 = 50$ is shown in Figure 6.5; this corresponds to a SNR of 6 dB. The samples were obtained by

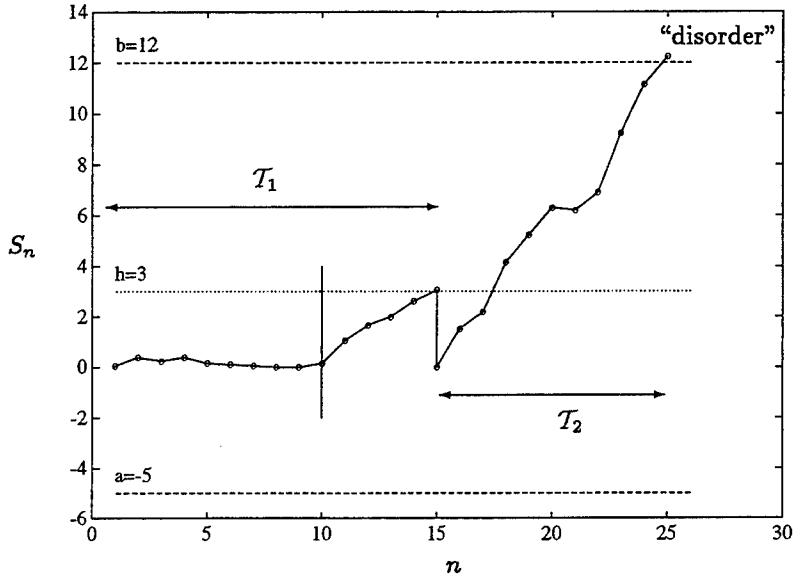

Figure 6.2: A sample realization of the composite detector when a disorder occurs. Here $\theta_1 = 15, \theta_{min} = 11,$ and $\theta_0 = 10$. The vertical bar indicates the true disorder time.

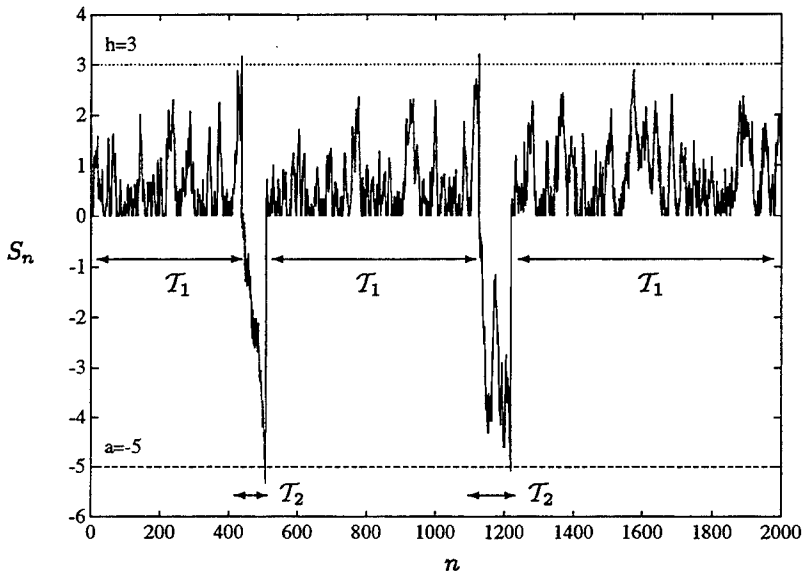

Figure 6.3: A sample realization of the composite detector when no disorder occurs. Here $\theta_{min} = 11$ and $\theta_0 = 10$.

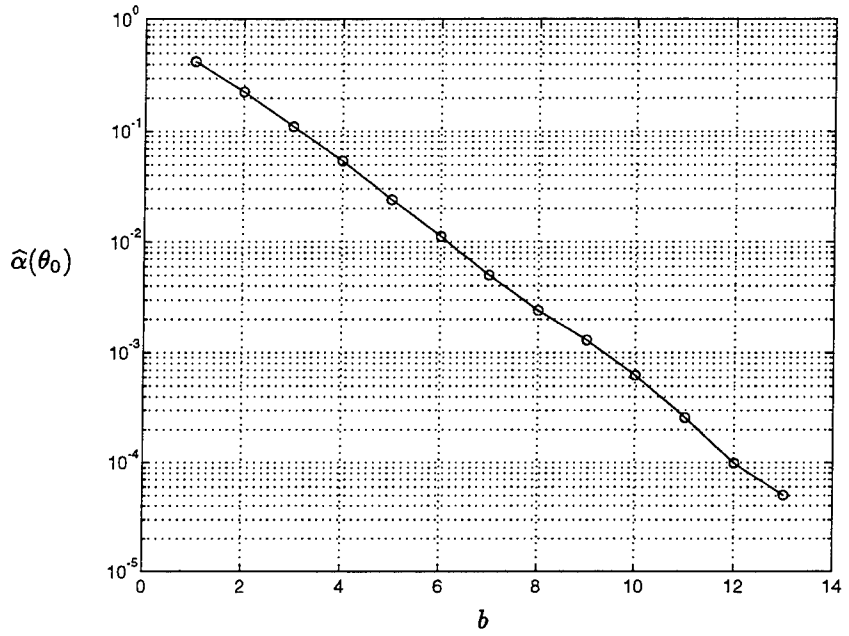

Figure 6.4: $\widehat{\alpha}(\theta_0)$ versus b.

letting b take on uniformly spaced values from 1 to 13. We see that the composite detector with $h = 6$ has both a greater expected delay (ED) and mean time between false alarms (MFA) than that for $h = 3$, for a fixed b. This occurs since a higher threshold in test $\mathcal{T}_1$ means more samples will be required to cause an alarm. We also note that the slope of the performance curves for the composite test and that of the quickest detector are similar. This results from the fact mentioned in Section 6.4 that the ECSPRT behaves (asymptotically) like the optimal SPRT under H_1 and like the "minimum jump SPRT" under H_0. From these findings, one might be tempted to make $h \approx 0$ (i.e., do away with $\mathcal{T}_1$ completely). However, this would increase the likelihood that the disorder will occur during test $\mathcal{T}_2$, which also increases the likelihood that $\widetilde{\theta}_n$ will be based on samples under both f_0 and f_1, an undesirable condition as discussed in Section 6.4. In the present analysis, the MFA and ED are computed assuming that the disorder occurs when test $\mathcal{T}_1$ is active, a good assumption

when h is not close to zero. The more general case where the disorder may occur when either test is active is left for future study.

Figure 6.6 illustrates the case where $\theta_1 = 20$, for an SNR of 0 dB. For both this case and that of Figure 6.5, observe that the composite test outperforms Hinkley's test for higher MFA; however, the particular MFA at which this occurs depends upon the choice of parameters and θ_1. This fact suggests that the decision to use the composite test or Hinkley's test depends on the desired MFA. Again notice that the slope of the composite performance curve is approximately the same as that of the quickest detector, but that there is an offset of a few samples between the two curves. This offset arises from two factors. First, the minimum number of samples required for the composite test is two, instead of only one for the quickest detector since the former is composed of two tests. Second, an additional delay is incurred in the former due to the additional time required for $\tilde{\theta}_n$ to converge "close enough" to θ_1 to allow the ECSPRT to react in a similar manner as the quickest detector.

Finally, Figure 6.7 shows the case where $\theta_1 = 11$, for a SNR of -10 dB. This is the case where the true jump is that of the minimum assumed magnitude. Here, the quickest detector and Hinkley procedure are the same test, and so only one performance curve is shown for both. The performance curves of the composite test are similar to that of the quickest detector. In particular, the slopes of the former appear to be only slightly greater than the latter. This is not surprising, since the optimal performance is achieved by using the quickest detector. We can also observe that the offset between the two curves becomes less critical in that the percentage increase in ED of the composite test over the quickest detector is smaller for smaller θ_1.

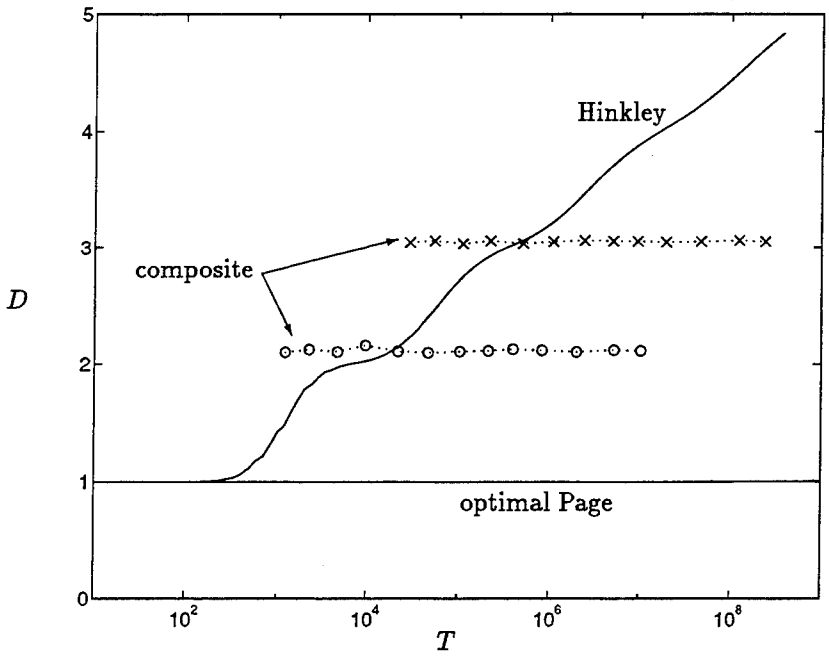

Figure 6.5: Performance of the quickest, Hinkley, and composite detectors for $\theta_1 = 50$ (SNR $= 6$ dB). For the composite detector, $a = -5$, $h = 3$ (○'s) and 6 (×'s).

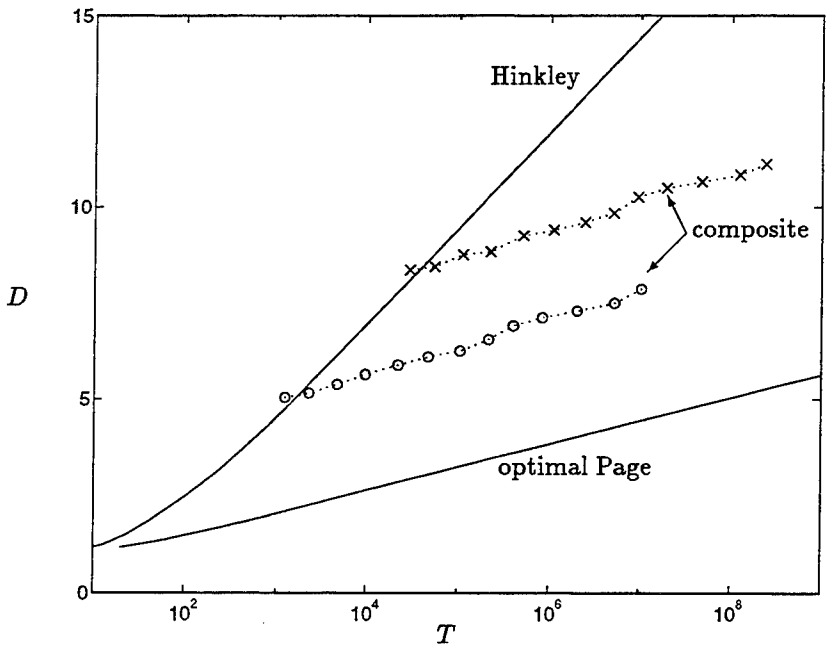

Figure 6.6: Performance of the quickest, Hinkley, and composite detectors for $\theta_1 = 20$ (SNR $= 0$ dB). For the composite detector, $a = -5$, $h = 3$ (○'s) and 6 (×'s).

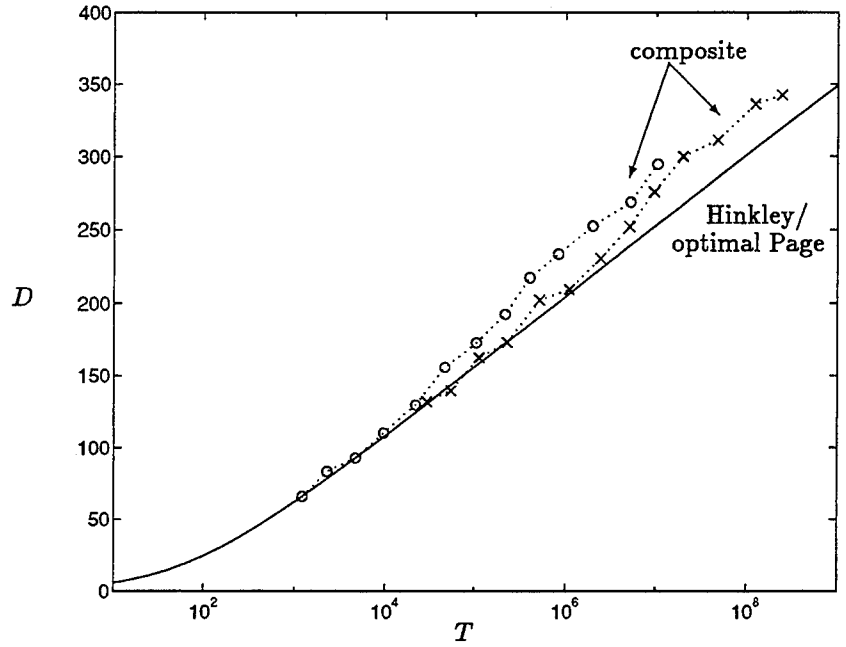

Figure 6.7: Performance of the quickest/Hinkley and composite detectors for $\theta_1 = 11$ (SNR $= -10$ dB). For the composite detector, $a = -5$, $h = 3$ (○'s) and 6 (×'s).

6.6 Conclusions

In this chapter, an adaptive approach for detecting a jump change of unknown magnitude in the parameter of a random process was introduced. Our approach, termed the composite test, is a heuristic procedure created with the goal of retaining the beneficial properties of the optimal Page's test, the quickest detector, even though the jump magnitude is not known. Our examples focused on the case where the disorder is a jump in the rate parameter of a Poisson process, but, as discussed in Section 6.4, the procedure can be applied to other processes as well.

In Section 6.5, the composite test was analyzed. It was shown that the performance can be expressed in terms of the performance of each of the sub–tests $\mathcal{T}_1$ and $\mathcal{T}_2$. Because closed form expressions for the ECSPRT were not available, Monte Carlo simulations were used instead.

The performance of the composite test relative to the quickest detector and Hinkley's test, which is Page's test designed for the jump of minimum magnitude, was evaluated using several examples. It was shown that the composite test outperforms Hinkley's test for higher mean time between false alarms. It also exhibits performance that is similar to the quickest detector in the sense that the slopes of the composite and quickest detectors are similar; therefore, the performance of these procedures will be asymptotically similar, differing by a bias which is proportional to the true jump magnitude. This work shows that the composite test is a viable procedure for detecting jumps in the mean of unknown magnitude.

There are several directions for future work. First, it would not be difficult to evaluate the performance of the composite test for other distributions, and also for the case where the samples are correlated. Second, it would be nice to obtain closed form expressions for the performance of the ECSPRT, which would allow us to express the performance of the composite test without the use of simulations. Finally, it would

be useful to develop an extension of the procedure to the multivariate case.

6.7 References

[1] M. Basseville, B. Espiau, and J. Gasnier, "Edge detection using sequential methods for change in level - Part I: A Sequential Edge Detection Algorithm," *IEEE Trans. Acous., Speech, Signal Processing*, vol. ASSP-29, no. 1, pp. 24-31, Feb. 1981.

[2] M. Basseville, "Edge detection using sequential methods for change in level - Part II: Sequential detection of change in mean," *IEEE Trans. Acous., Speech, Signal Processing*, vol. ASSP-29, no. 1, pp. 32-50, Feb. 1981.

[3] M. Basseville, "On-line detection of jumps in mean," in *Detection of Abrupt Changes in Signals and Dynamical Systems*, M. Basseville and A. Benveniste, ed., Springer-Verlag, New York, 1986, pp. 11-26.

[4] B. Broder and S. C. Schwartz, "The performance of Page's test and nonparametric quickest detection," *Proc. 1989 Conf. Information Sciences and Systems*, Johns Hopkins University, Baltimore, MD, pp. 34-39.

[5] R. W. Crow and S. C. Schwartz, "An adaptive procedure for detecting jump changes in Poisson observables," *Proc. 1993 Conf. Information Sciences and Systems*, Johns Hopkins University, Baltimore, MD, pp. 84-89.

[6] P. Papantoni-Kazakos, "Algorithms for monitoring changes in quality of communication links," *IEEE Trans. Commun.*, vol. COM-27, no. 4, pp. 682-693, Apr. 1979.

[7] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*, Springer-Verlag, New York, 1985.

[8] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, John Wiley and Sons, Inc., New York, 1968.

Chapter 7

Conclusions

This thesis investigated quickest detection procedures for several types of disorder problems. The objective throughout was to minimize the expected delay in detecting the disorder, subject to a lower bound on the mean time between false alarms. The contributions of this thesis are given below.

7.1 Contributions

Chapter 2 provided an overview of the foundations of work on quickest detection, and served as a prerequisite for the rest of the thesis. The asymptotic performance measure (APM) was shown to be a useful figure of merit, in that it allows us to characterize the performance of a detection procedure using a single number. Consequently, the asymptotically optimal procedure can be determined via the maximization of the APM. Since the computation of the APM is not always feasible, a lower bound that approximates the APM was used in the design process. It was shown that the log-likelihood ratio is necessary and sufficient to maximize this bound. Several techniques for computing the performance of procedures to detect disorders were also given.

Quickest detection procedures when underlying noise models were partially unknown were considered in Chapter 3. Specifically, the ε-contamination and total

variation uncertainty classes were studied. The minimax robust quickest detector was derived by applying the minimax criterion directly to the APM, and it was shown that a saddlepoint solution exists for this problem. The minimax APM was shown to equal the Kullback-Leibler (K-L) divergence, and so the least favorable distributions are those which minimize this quantity; the robust processor is then the log-likelihood ratio of the least favorable densities. Performance curves are generated which show that the robust procedure works well for a number of members of the uncertainty class, and in all but a few cases outperforms nonparametric techniques based on the linear, sign, and dead-zone nonlinearities. For the weak signal case, we established an equivalence between the APM, the classical efficacy, and Fisher's information. The weak signal robust detector was obtained by finding the least favorable distribution for Fisher's information. Performance curves were given to show the gains available when robustness is built into the detection procedure.

The investigation of robust quickest detection procedures was continued in Chapter 4 for the case where the mean and/or noise covariance of a multivariate Gaussian process is uncertain. As in the previous chapter, the robust procedure is obtained by applying the minimax criterion to the APM. It is shown that the robust processor is exactly the robust discrete-time matched filter. Particular solutions were presented for several different uncertainty classes, each of which is based on the deviation from some nominal parameter set. Some performance curves were given which illustrate the tradeoffs when there is a mismatch between the assumed and actual levels of uncertainty. The applicability of the robust procedure to non-Gaussian multivariate processes was also discussed.

In Chapter 5, we examined the problem of designing quickest detection procedures at the fusion center of a distributed detection system. An optimal procedure was derived and compared to several alternative methods which are easier to implement in

that they are recursive and require less computation. It was shown that a slight modification of the optimal scheme leads to a suboptimal procedure whose performance differs negligibly from the optimal. A simple method for choosing the thresholds of the local detectors was presented; specifically, the thresholds were selected to maximize the asymptotic performance measure of the distributed system. A sensitivity analysis revealed that the method results in overall system performance which is close to optimal, even for small mean times between false alarms. Lastly, the relationship between channel bandwidth and detection delay was evaluated. It was shown that the optimum bandwidth is a function of signal strength. Perhaps contrary to intuition, for weaker signals, the optimal performance did not result from the system with the maximum bandwidth. Performance curves were presented which illustrate the performance gain or loss as the bandwidth varies; these curves would be useful to a designer who must make decisions based on the tradeoff between bandwidth cost of bandwidth and system performance.

Finally, the disorder problem when a jump of unknown magnitude occurs in the mean of a random process was investigated in Chapter 6. Optimal methods exist when the jump magnitude is known (Page's test using the log-likelihood ratio); we proposed an adaptive procedure suitable when this information is not available. The procedure consisted of two tests which operate in series. The first test signals a candidate disorder time, which the second test then uses to form an estimate of the post-disorder mean value. This estimate was then incorporated into an adaptive version of the well-known sequential probability ratio test. The average sample number (ASN) of the adaptive procedure was derived and expressed in terms of the two sub-tests. It was shown via simulation that the adaptive test has similar asymptotic performance to the test which is optimal for known jump size. The procedure was implemented to detect a change in the rate parameter of a Poisson process; however, it is also applicable to other distributions.